Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 146 (2019) 201–207

Procedia Computer Science

www.elsevier.com/locate/procedia

14th International Conference on Current Research Information Systems, CRIS2018

Entanglement of bibliographic database content and data collection practices: Rethinking data integration using findings from a European study

Linda Sīle*

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

Abstract

This paper proposes *transparency* and *reflexivity* as two principles to be incorporated in initiatives wherein data from multiple national contexts are integrated. The necessity of these principles is derived from an ongoing study tasked with identifying and describing national bibliographic databases for research output within the social sciences and humanities (SSH) in Europe. The study is carried out within the context of the COST Action "European Network for Research Evaluation in Social Sciences and Humanities" (ENRESSH). Within ENRESSH, it is emphasised that national bibliographic databases can be instrumental in enhancing the visibility of research within SSH. Hence, one of the aims of ENRESSH is to identify and describe currently existing databases and eventually design a roadmap for a European database that would include data on research output within SSH from different European countries. The study shows that there are considerable challenges in merely acquiring a basic description of the content of databases embedded in different national contexts. To make sense of the content it is necessary to acknowledge the role of context in information systems. Emphasising context, as will be shown, it is possible to elucidate the nature of the encountered challenges as well as to highlight aspects to be incorporated in designs of research information systems.

Keywords: research information system; database; research output; social sciences and humanities; Europe; data integration.

* Corresponding author. *E-mail address:* Linda.Sile@uantwerpen.be

10.1016/j.procs.2019.01.094

1. Introduction

In recent decades, a number of countries and regions have implemented national bibliographic databases either as modules within current research information systems (RIS) or as separate databases on the national or regional level (e.g., VABB-SHW in Flanders, Belgium, RIV in the Czech Republic). One of the goals of these initiatives is to acquire comprehensive coverage of national research output within the social sciences and humanities (SSH). A characteristic of SSH is the great diversity in media which are used to communicate or to represent research findings. Research output in SSH is relatively less visible in commercial international databases such as the ones in Web of Science (www.webofscience.com) or Scopus (www.scopus.com). Consequently, for SSH, research monitoring, assessment and/or funding allocation that incorporates bibliometric indicators is rather problematic given the absence of data that represent the rich variety of scholarship within this knowledge domain. For this reason, there is a particular value in RIS with comprehensive data on research output within SSH.

It can be problematic to identify and negotiate relevant categories to be incorporated in RIS designs, especially when an envisioned RIS spans multiple contexts (e.g. institutional, regional, or national)[1]. In such a situation, in addition to the identification of relevant categories, one has to estimate to what extent different categories are comparable across different contexts and, how one ought to proceed in cases when incompatible categories are used in different contexts. Aside from these content-oriented considerations, there is also the pragmatic side: time can be a limitation when it comes to an exploration of the different contexts. Similarly, there can be limited access to the necessary information (either due to a language barrier, differences in professional background, or else). Hence the key problem addressed in this paper is a reconciliation of these two aspects—awareness of contextual aspects and the need to integrate data—when integrating data from different national contexts.

The need for comprehensive data on research output in SSH is acknowledged within the COST Action "European Network for Research Evaluation in Social Sciences and Humanities" (ENRESSH, www.enressh.eu), a network launched in 2016. In ENRESSH, it is emphasised that national bibliographic databases (here treated as basic RIS) can be instrumental in enhancing the visibility of research within SSH. Guided by this rationale, the work within ENRESSH is carried out towards a roadmap for a European database for research output within SSH, created by integrating data from RIS that exist across the countries in Europe. The first step towards this end is a study of currently existing databases in Europe, the findings of which form the empirical material underpinning this paper.

This paper is a conceptual contribution to principles behind integration of data on research output in SSH drawn from databases in different national contexts. Specifically, *transparency* and *reflexivity* are proposed as two principles to be incorporated in data integration initiatives. The necessity of the two is derived from an ongoing study on national bibliographic databases for research output within SSH in Europe. This study has shown that there are considerable challenges in merely acquiring a basic description of the content of the different databases. To make sense of the content as well as of the encountered challenges, it is useful to incorporate insights from literature that foregrounds the role of contexts in information systems[2,3,4,5]. These insights help to, as will be shown, to elucidate the nature of the encountered problems and also points to directions for novel ways of thinking about designs of RIS that integrate data from different national databases for SSH research output.

The structure of this paper is as follows: first, I highlight the need to explore data collection and processing practices when attempting to capture the kind of content that is present in national bibliographic databases. To illustrate this, I continue with a brief description of a study the aim of which was to identify and describe currently existing national bibliographic databases for research output in the social sciences and humanities. Then, I bring to attention challenges that were encountered in the process of describing the different databases. In the final section, follows a conceptualisation of the encountered challenges; this conceptualisation underlies a proposal of two principles—*transparency* and *reflexivity*—for data integration initiatives in cases when data are drawn from multiple national contexts.

2. Making sense of the content of national bibliographic databases for research output

The content of databases for research output seems to be a straightforward matter. Such databases collect (metadata on) research output. Given the increasing popularity of institutionalised research output metadata collection activities, it should perhaps be self-evident how to describe the content of databases of this type. We could use research output

types and, perhaps, publication years as properties to characterise the database content. However, experience from a recent study aiming to accomplish such a descriptive task shows that for understanding of databases content there is a need to take into account that database content is tightly entangled with data collection and processing practices. To illustrate this, I continue with challenges encountered in a study of European databases for research output in the social sciences and humanities.

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

The aim of the study was to identify and describe currently existing national bibliographic databases for research output in the social sciences and humanities. The study consisted of two surveys coupled with a participant observation of databases being surveyed. The aim of the first survey was to identify and briefly describe currently existing databases (scope: 41 countries; 95% response rate). The second survey was focused on the comprehensiveness and data processing of a selection of national databases (scope: 17 countries; 76% response rate). The purpose of the participant observation was to acquire more detailed information on databases which may have fallen beyond the sensibilities of the two surveys as well as to enable a reflection upon the extent to which the insights inquired through surveys were determined by assumptions built into the survey designs.

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 or 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, appears the most appropriate in the first attempt to systematically collect information on currently existing national bibliographic databases for SSH research output in Europe.

The first stage of the study resulted in responses from 39 countries (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). The key finding of the first survey is that there are 21 national databases in 20 European countries. Databases were identified in 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. In Albania, Latvia, and Portugal, as well as in Serbia new databases are currently being set up. At the same time it is also known that there are still more databases. In some cases participants indicated that there is more than one national database; in most such cases, only one database was described.

In terms of the most common types of publications included in national bibliographic databases, the only publication type included in all 21 databases is the journal article. The majority of databases, though, also include data on monographs (n=17), edited volumes (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 the 1960s, while in the databases in, for example, the Czech Republic, Estonia, and Slovenia one can find data on research output from 1990s. It is not, however, known at the moment how comprehensive are the records going further back in history. More detailed results from the two surveys can be found elsewhere[6,7].

2.2. Challenges in making sense of the database content

The brief summary of the most common publication types included in national bibliographic databases in the previous section is presented as if it is equivalent across all databases. This summary, even though informative, does not contain any practice-related features that, in some cases, may be essential to understand how (if at all) data can be successfully integrated.

We do know that there is also variation 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 other staff) is combined with data transfer from research organisations, publishers and other national or international databases (e.g. Web of Science and/or Scopus). This information, though insightful, is still insufficient since it is not clear what are the implications from the different data collection methods for the database content. We do not know which methods lead to more accurate metadata. We do not know whether data transfer in each of the databases is carried out along the same principles. These aspects, however, are crucial when designing a data integration project and trying to understand which entities and their properties can be deemed equivalent.

Illustrative of this is experience from the study of databases. In the first survey, information on the included output types was acquired asking the survey respondents to mark whether a national database includes, for example, 'book chapters' (no definition was used). In the second survey, the category 'book chapters' was named 'articles or chapters in books' denoting an "independent part of a monographic publication or an editorial collection (excluding articles in conference proceedings)". The hope was that the overview of output types acquired using category names in combination with definitions will be more accurate and valid. Definitions were provided in the questionnaire, and, in addition, more detailed remarks were provided in a manual provided to the study participants. For each output type, participants were invited to indicate whether there is such a category in a database and if so, whether the definition we provide corresponds to the definition that is (implicitly or explicitly) employed in a database. In cases of mismatch between definitions, we encouraged participants to describe in what way the definitions differ. Taking a closer look into data collection practices it became evident that also this approach is not sufficient. Even though formally this definition seemed to be acceptable across the different databases, the actual practices turned out to rely on contingent features which fall beyond both the definitions and the generic descriptions of approach to data collection.

To give an example, I contrast two databases: the Flemish database VABB-SHW and the Norwegian Science Index (NVI), a subset of the bibliographic database within the Norwegian RIS Cristin. From the study by Kulczycki and colleagues[8] we know that for Flanders the share of book chapters is twice the share found for Norway. However, we do not know whether the category 'book chapter' can be treated as equivalent across the two databases. The two databases are relatively similar. The content in both databases is peer-reviewed scholarly publications restricted to a selection of publication types (VABB-SHW: monographs, edited volumes, book chapters, contributions in conference proceedings, and journal articles; NVI: monographs, book chapters, and journal articles). A key difference concerning book chapters is related to another category—contributions in conference proceedings. In VABB-SHW publications of this type are collected in a designated category. In contrast, in NVI, contributions in conference proceedings are classified as book chapters or journal articles depending on whether the parent publication is a journal (with ISSN) or a book (with ISBN). To acquire consistent data, the easiest approach would be to reclassify VABB-SHW records within the category 'contributions in conference proceedings' (those with ISSN: to journal articles; those with ISBN: to book chapters). The inverse process (re-classifying NVI categories 'book chapter' and 'journal article') would be more time consuming. Furthermore, it is quite telling that the definition for the category 'book chapters' we used did specify that the category book chapter should not include contributions in conference proceedings. Concerning NVI, no additional remarks were provided thus implying that the category does match the definition.

An even more subtle discrepancy stems from general inclusion criteria guiding data collection in the two systems. In both, a general definition of publications is used to delineate the scope of the database. Both definitions contain a requirement of scholarly content. In NVI, a scholarly publication must "present new insight"[9], but in VABB-SHW, a publication must "make a contribution to the development of new insights or to applications resulting from these insights"[10]. In practice, the two requirements are translated in slightly different way. In NVI, this requirement means that introductions, prefaces, conclusions or other similar chapters are not reported, unless a special argument is made that a chapter of this type does contain scholarly content and is presented in a scholarly form. In VABB-SHW, in contrast, such a differentiation is not carried out systematically (normally, prefaces and forewords are not included; introductions are). Instead, there is a requirement for publications to be at least four pages in length. Also here it is evident that if data from such databases are integrated, then the result concerning the category 'book chapters' is no longer consistent. An attempt to incorporate the same principle for the two databases would require, first of all, a choice of a specific standard, and, second, additional effort amount of time in case any of the records would require reclassification. This is another example of the entanglement of the content and data collection and processing

practices. It is tempting to think that these examples are exceptions. The study however shows that there are numerous aspects in relation to which the content of databases differs[7].

3. Rethinking data integration

3.1. RIS as infrastructure

The examples of the more or less subtle ways of the entanglement between data content and data collection practices bring to attention a more general reasoning beneath our thinking about bibliographic databases, RIS, and data integration. To elucidate this, I draw on conceptual insights from infrastructure studies and specifically the work by Susan Leigh Star and her colleagues[2,3,4]. Star and Ruhleder[4] highlight the relational nature of infrastructure. We speak of infrastructure as a constellation of objects, people, and practices *on which* other tasks are carried out. There is no general term to define the *kind* of entity that infrastructure is precisely because the very essence of infrastructure resides not within itself, but within the practices the infrastructure supports. Transferring this conceptualisation to RIS it becomes evident that without the ideas of research information, of research monitoring, auditing and the numerous reporting procedures reliant on data within RIS, RIS is merely a container (for a selection) of data. The meaning of the data contained as well as of the rationale guiding the design of RIS, resides in the activities for the support of which RIS was created. In addition, we must not forget also the people, each with their own work conventions, tasked with data input, data retrieval, development of parsers and other handy algorithms, and the many other tasks which enable smooth operation of such systems.

One of the implications from this conceptualisation of RIS is that it allows to reconsider the relation between data collection and processing practices and the RIS content. It is commonplace to think of variations that one encounters in RIS work practice as a combination of an ideal practice and human error. I shall note that by RIS work I mean all work practices that underpin existence of RIS (starting from design phase, through implementation, and maintenance). Seeing RIS work in these terms, the improvement of RIS requires ever better notions of ideal practice and ever less human error. In other words, the focus is on standards: standards in theory, and standards in practice. In contrast, if RIS is conceptualised as infrastructure in relational terms, it becomes evident that, firstly, RIS are not value-neutral. As a range of studies that conceptualise information systems in similar terms have shown[2,3,4,11,12], information systems carry social structures and hierarchies, values and conventions along with all the other characteristics of contexts in which systems are situated. Certain understanding and priorities or simply ‘ways of doing’ are folded into infrastructures. Acknowledging this aspect, variation in practices emerges as an indication of unique social practices each with its (more or less) distinct features which can be explained as a result of differing value-orientations, conventions or other contextual factors (and not always as human error).

These differing value-orientations and conventions have been noted also in data integration projects that take place in one national context[1,13]. Yet, when it comes to multiple national contexts, these issues become even more apparent as illustrated with the example of two databases and the category ‘book chapter’. In more general terms in relation to research output in SSH, these aspects can be potentially crucial since it is not rare for certain SSH disciplines to have forms of communication characteristic to one specific national (or regional) context. Hence for integration of data on SSH research output drawn from multiple national contexts two problematic aspects emerge: the problem of identification of differences in practice and the problem of choosing standards.

3.1.1. The problem of identification and boundary objects

To capture discrepancies in practices, an especially useful concept is ‘boundary object’[3]. Boundary objects, as developed by Star and Griesemer, are abstract or concrete entities that reside in different social worlds and are flexible enough to allow across-worlds communication and collaboration without losing its meaning. Typically, an entity fixed with a specific name and a detailed enough definition is assumed to be sufficient to model a representation of a specific social phenomenon in an information system. Using the notion of ‘boundary objects’, such an assumption is turned into a hypothesis that can be explored by investigating the ways how activities related to a specific term are carried out in practice.

These rather abstract considerations played out in the database surveying process as an ever growing awareness that the use of equivalent definitions is still insufficient to capture the content of databases. And this is because not only the very name of the category, but *the combination of the name and its definition* act as boundary objects. By specifying a category and a definition, one acquires means to communicate and coordinate social (inter-)action while the content of a category remains elusive and situated in practices that are still unknown. It is not to say that it is impossible to capture the database content and introduce more refined definitions. Quite the contrary, the content can be captured and suitable definitions can be phrased. However, such an achievement requires a theoretical framework as well as methods that are sensitive to interpretive subtleties all along the way from a record in a database back to the contexts from which it originates.

3.1.2. The problem of choosing standards and performativity

A further conceptual implication is that databases, information systems in general and hence also RIS are *performative*[2,11,12]. Performativity here refers to a capacity to reinforce a specific understanding of a phenomenon by embedding it in the design of RIS. In other words, RIS carry strong epistemic authority which turns out problematic in the presence of multiple standards to choose from. If the problem of identification is overcome and the origins of data to be integrated are traced, then next, one is faced with the problem of choosing standards. When different standards are due to cultural conventions, traditions and simply different value-orientations it is no longer evident which standard to choose from. The earlier provided example of differences in the content of a category 'book chapter' in two databases highlights two different practices which can be treated as two standards. If such data would be integrated, for the consistency of data in this category, one could choose from the following options:

1. To integrate data without any alterations in data;
2. To integrate only that subset of data using an intersection *both* standards;
3. To integrate only that subset of data using *one* of the two standards.

The problematic side of this choice pertains to consequences for bibliometric indicators produced using such integrated data. There can be a risk to either acquire insufficiently robust indicators (due to inconsistent data; option 1) or to over- or underestimate, for example, the volume of research output (options 2 and 3). Similarly, at the point when a new RIS is implemented, all the ways it will be used are not known. For this reason, it is not known which design features will turn out central (and perhaps problematic) for some users or those to whom the data in RIS refer.

4. Reflexivity and transparency in data integration

There are many unknowns in data integration. It is not evident how detailed information about data to be integrated is sufficient. It is not clear what consequences (if any) for research may follow from continuous auditing and monitoring of research. In the same way, consequences from the choice of one standard are typically not known at the point when choice is being made. And finally, regardless of the broader epistemic challenges, data integration initiatives may simply be with practical constraints for, for example, time that can be devoted to database explorations and reflections on performativity. For these reasons, it seems useful to think of more general considerations that could guide the designs of data integration projects that involve data produced in differing national contexts. Here, I propose two such possible considerations: transparency and reflexivity.

'Transparency' here refers to openness about, here, data integration process. It somewhat mirrors the reusability principle in FAIR. To integrate data, we need to know what phenomena the data represent. We need to know their origin, the content, and the way the data were created and perhaps altered. Such claims might seem as stating the obvious: who would doubt that the content of RIS depends on data collection and processing methods? At the same time, the experience with the database surveying process shows that the practices that underlie database setups are more varied than anticipated. Moreover, in some situations, specifics become apparent only after one is familiar enough with a specific context and knows what questions to ask. For this reason, it is useful to keep in mind that information in RIS is a representation that results from a sequence of actions (by humans or technologies), each of which are coupled with considerations that affect the meaning of the RIS content to a smaller or greater extent. The more practical consequence from this is that data integration requires, first, an inquiry, sensitive to multiple interpretations, of the data sources to be integrated. Second, this transparency principle also requires sufficiently

detailed documentation of considerations guiding, for example, the mapping of different classification schemes, and perhaps, most importantly, an explication of moments when a specific data integration design decision is made without knowing practices through which data to be integrated went through.

The other, the reflexivity principle is closely related and refers to awareness of the performativity of RIS. This is a more future and consequence-oriented principle which calls for more attention to social consequences that may follow to certain RIS design decisions. As explained earlier, as soon as user interfaces are implemented or data reporting is launched, details of considerations that went into the creation of RIS tend to be left behind. This, in principle, is the way how one might expect smooth operation of RIS. At the same time, there is a risk that without reflexivity on behalf of RIS designers, it can be forgotten that different design decisions could have been made and the representation of research based on RIS data could have looked otherwise. Similarly, without explication of limitations of RIS, there is a possibility that data are used for purposes for which they are not suitable. For this reason, it is useful to introduce more reflexivity in the work around RIS. Much work has been done in infrastructure studies in showing how certain information systems go better along with certain practices, cultures, and values[5,11,12]. Therefore, there is a scope for ways to design RIS that not only decrease the administrative burden for different organisations (an often used rationale[1,13]), but also to take into account the performative role that RIS play in research system and think of features that allow for more inclusive data collection approaches and/or openness to multiple standards. How these ideas can be translated in more practical terms, requires interdisciplinary collaboration wherein insights from social theory are combined with expertise in the design, implementation, and integration of information systems.

In this paper I have restricted the discussion to national bibliographic databases for research output and RIS. However, the same reasoning is applicable to any information systems that aim to represent artefacts that span multiple social, cultural, and temporal contexts.

Acknowledgements

The author thanks Tim Engels, Raf Guns and reviewers for the valuable feedback on ideas presented here.

References

1. Biesenbender S, Hornbostel S. The Research Core Dataset for the German science system: challenges, processes and principles of a contested standardization project. *Scientometrics* 2000, 106(2), 837–847.
2. Bowker GC. *Memory practices in the sciences* (1. paperback ed). Cambridge, Mass.: MIT; 2008.
3. Star SL, Griesemer JR. Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, 1989. 19(3), 387–420. https://doi.org/10.1177/030631289019003001
4. Star SL, Ruhleder K. Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research* 1996, 7(1), 111–134.
5. Waterton C. Experimenting with the Archive: STS-ers As Analysts and Co-constructors of Databases and Other Archival Forms. *Science, Technology & Human Values* 2010, 35(5), 645–676.
6. Sīle L, Guns R, Sivertsen G, Engels TCE. *European Databases and Repositories for Social Sciences and Humanities Research Output* (p. 25). Antwerp: ECOOM & ENRESSH; 2017. Retrieved from https://doi.org/10.6084/m9.figshare.5172322.v2
7. Sīle L, Pölönen J, Sivertsen G, Guns R, Engels TCE, Arefiev P, … Teitelbaum R. Comprehensiveness of national bibliographic databases for social sciences and humanities: findings from a European survey. *Research Evaluation* 2018. rvy016 https://doi.org/10.1093/reseval/rvy016
8. Kulczycki E, Engels TCE, Pölönen J, Bruun K, Dušková M, Guns R, … Zuccala A. Publication patterns in the social sciences and humanities: evidence from eight European countries. *Scientometrics* 2018. 116(1), 463-486.
9. Sivertsen G. Publication-Based Funding: The Norwegian Model. In M. Ochsner, S. E. Hug, & H.-D. Daniel (Eds.), *Research Assessment in the Humanities*. Cham: Springer International Publishing; 2016. p. 79–90.
10. Verleysen FT, Ghesquière P, Engels TCE. The objectives, design and selection process of the Flemish Academic Bibliographic Database for the Social Sciences and Humanities (VABB-SHW). In W. Blockmans, L. Engwall, & D. Weaire, *Bibliometrics: Use and Abuse in the Review of Research Performance*. London: Portland Press; 2014. p. 115-125.
11. Bowker GC, Star SL. *Sorting things out: classification and its consequences* (First paperback edition). Cambridge, Massachusetts London, England: The MIT Press; 2000.
12. Lampland M, Star, SL (Eds.). *Standards and their stories: how quantifying, classifying, and formalizing practices shape everyday life*. Ithaca: Cornell University Press; 2009.
13. Vancauwenbergh S. Governance of Research Information and Classifications, Key Assets to Interoperability of CRIS Systems in Inter-organizational Contexts. *Procedia Computer Science* 2017, 106, 335–342.