Integration of national publication databases – towards a high-quality and comprehensive information base on scholarly publications in Europe

Hanna-Mari Puuska[1,*], *Joonas* Nikkanen[1], *Tim* Engels[2], *Raf* Guns[2], *Dragan* Ivanović[3], and *Janne* Pölönen[4]

[1] CSC – IT Center for Science Ltd., P.O. Box 405, 02101 Espoo, Finland
[2] Centre for R&D Monitoring (ECOOM), Faculty of Social Sciences, University of Antwerp, Middelheimlaan 1, B-2020 Antwerp, Belgium
[3] University of Novi Sad, Dr Zorana Djindjica 1, 21000 Novi Sad, the Republic of Serbia
[4] Federation of Finnish Learned Societies, Snellmaninkatu 13, 00170 Helsinki, Finland

Abstract. The need for a comprehensive infrastructure for scholarly publication information has been on the EU's agenda for a long time. Also, the European Commission's open science policy highlights the necessity of a good information base to follow up open access publishing across Europe. However, an all-inclusive information infrastructure on research publications is still missing. During the past 10 years, European countries have invested significantly in national research information infrastructures. Now, at least 20 European countries have a national database for research publication metadata. The strength of these databases lies in their comprehensiveness and quality assurance since they often have a mandatory nature. They are, however, neither yet integrated nor widely used for cross-country comparisons. To this end, a proof of concept of a European publication infrastructure was carried out in the framework of ENRESSH (www.enressh.eu). The ENRESSH-VIRTA-PoC integrated publication data from four countries and the concept was built on the strengths of the Finnish national VIRTA system. This paper highlights the results from the PoC and outlines future steps towards the integration of national publication databases in Europe.

1 Background

Science policy at all levels in Europe needs reliable information on research activities to be able to evaluate research activities, their quality, and the development of open science. The issue of a comprehensive infrastructure for scholarly publication information has been on the agenda in various EU policy documents [1-4]. Also, the call for complete open access in EU member states [5] and the Plan S [6] initiative highlight the need for comprehensive and comparable information to follow up on the development of open access publishing across Europe.

A comprehensive information infrastructure on research publications, however, is still missing. The most widely used commercial publication databases, Web of Science and

* Corresponding author: hanna-mari.puuska@csc.fi

Scopus, suffer from significant lack of coverage especially in social sciences and humanities [7-8].

Michael Gusenbauer [9] created a comparative picture of 12 of the most commonly used academic search engines and bibliographic databases and found that Google Scholar, with 389 million records, is currently the most comprehensive academic search engine. The aggregating harvesters, such as Google Scholar and Microsoft Academic use well-developed algorithms to inclusively gather all available sources. They have wide coverage and they also include books and national publications widely [2]. However, because they do not disclose their sources, it is not known which publications are left out.

For purposes of exploring the open research literature, OpenAIRE provides an inclusive platform but it covers only a small share of the total publication output of research organizations since it harvests primarily freely available documents and metadata from institutional repositories. OpenAIRE is comprehensive in terms of scientific fields but its coverage is 'accidental' rather than systematic. For Google Scholar, Microsoft Academic, and OpenAIRE it is not possible to assess to what extent they are biased in terms of languages, publication types, open access, or scientific fields. Therefore, they are not suitable for research evaluation purposes.

Another limiting factor is that many services, such as Dimensions, rely on the availability of Digital Object Identifiers (DOI). This is also the case with the Unpaywall service, which can be used to identify different types of OA publications based on DOIs [10]. There are, however, considerable differences in DOI availability between different publication types, fields, and countries [11-13].

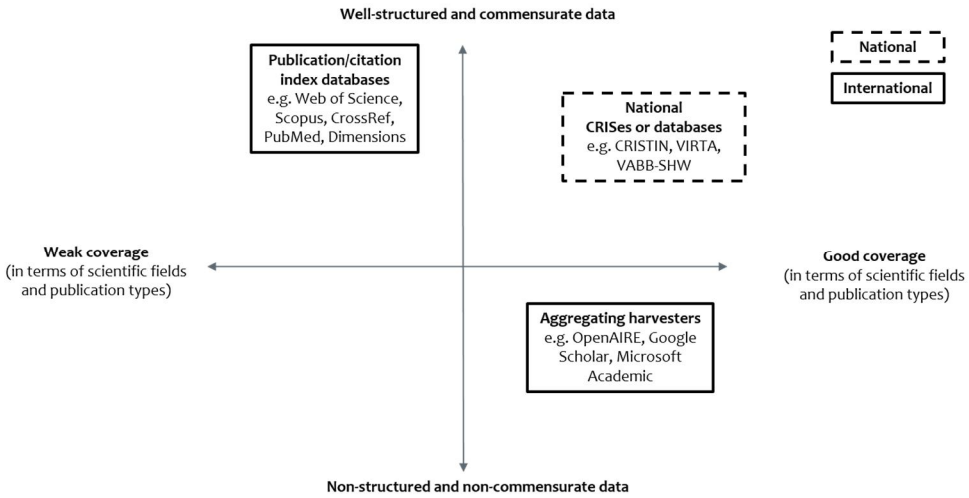

Fig. 1. Different types of publication databases.

During the past decade, European countries have heavily invested in national and regional research information infrastructures with the aim to better monitor and assess the research performance in their country. National publication databases have been implemented since the other existing databases do not give a comprehensive picture of publication activities (see Figure 1). These national databases are either 1) centralized information systems into which all organizations input metadata on their publication and other research output, or 2) harvest information from research organizations' own CRIS systems (Current Research Information Systems).

A recent survey found out that 20 European countries have a national database for research output [14]. Many of these databases cover all fields of science. A particular strength of the national databases lies in their comprehensiveness and high level of quality assurance since they often have a mandatory nature. For example, Denmark, Finland, Hungary, Italy, Norway, Poland, and Slovenia use their national publication database as a basis for universities' governmental performance-based research funding allocation. The databases are also used for national or organizational research assessments.

Several recent declarations call for responsibility in research evaluation. DORA - Declaration on Research Assessment [15], the Leiden Manifesto for research metrics [16] and the Metric Tide report [17], recommend that the research evaluation should be based primarily on expert judgment, but may be supported by metrics which presume robustness, transparency, diversity, and reflectivity of the data. The national publication databases correspond well for the needs of responsible metrics due to their very strict quality criteria and validation procedures. Moreover, they are transparent since the publication metadata is usually publicly available. Most of them have complete coverage of the country's peer-reviewed publications in all fields. In addition, they take into account the diversity of publishing by not just including journal articles but also conferences, monographs, and edited works in all languages.

The national databases enable comprehensive monitoring of different aspects of scholarly publishing: the development of the total volumes, the occurrence of different publication types, scientific fields, and languages, as well as the development of open access publishing. However, the national publication databases have not yet reached their full potential for cross-country comparisons. A few studies, in which national publication data from several European countries have been integrated manually, have demonstrated the great potential of these data sources for improved understanding, for example, of publication patterns and multilingualism in scholarly communication [18-19].

2 Towards a European infrastructure of publication information

The survey on national databases [14] indicates that the main challenge of the interoperability of publication information across countries is the variety of data models, vocabularies and classifications, as well as of data collection and validation procedures. Many countries have solved similar problems at the national level in different ways when compiling information from research organizations' local CRISes.

To this end, a proof of concept of a European publication infrastructure integrating national databases was set up in the framework of ENRESSH. The European Network on Research evaluation in Social Sciences and Humanities (www.enressh.eu) is an EU funded COST action network with partners from 36 European Countries. The network aims at advancing understanding and evaluating SSH research. One of the working groups of ENRESSH is specifically set to coordinate initiatives in terms of standardization and interoperability of research information in SSH and to design a roadmap for a European database for research outcomes.

The ENRESSH-VIRTA-PoC [20] was carried out in 2017 with publication data from Belgium, Finland, Norway, and Spain, building on the strengths of the Finnish national VIRTA Publication Information Service. In VIRTA, launched in 2016, the Finnish organizations store a copy of publication information of their institutional publication databases. The organizations use various local solutions from commercial CRIS systems to self-made publication registers. VIRTA is a data warehouse, "a data hub", making up-to-date metadata from research institutions available for other services and producing

comprehensive and comparative information on publishing activity both nationally and institutionally. Also, a simple Publication Input Service has been built to Finnish organizations that do not have their own CRIS system.

The proof of concept introduced a solution that integrated and validated metadata on research publications from different source systems (Figure 2). Since the national databases as source systems vary from simple databases to advanced CRIS systems different solutions should be provided for the data transfer. Both REST API and OAI-PMH endpoints, as well as XML file transfer, are supported in the suggested technical solution. Potentially also a publication input service would be needed for organizations or countries without a proper publication database.

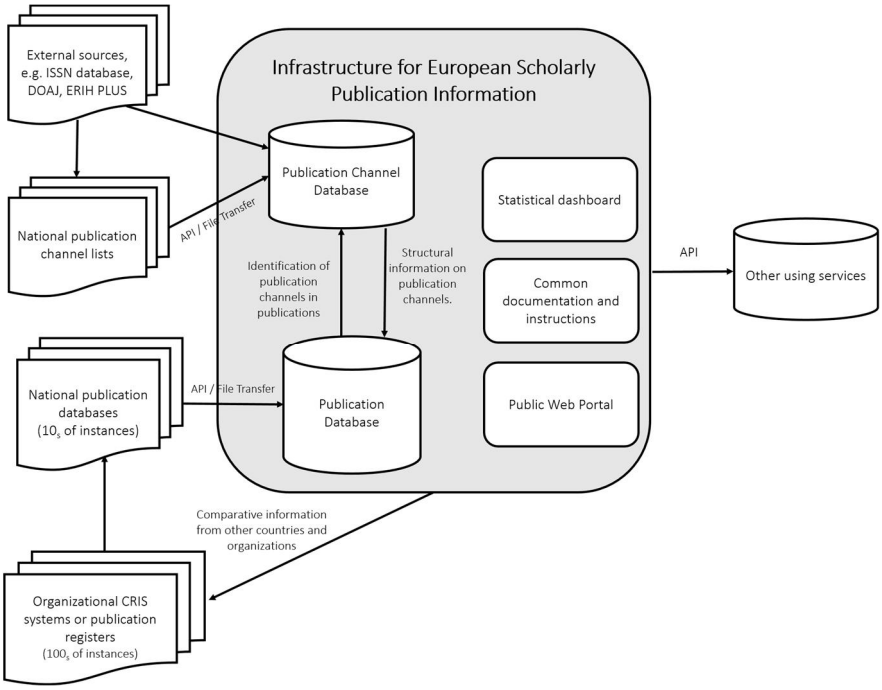

Fig. 2. Outline of the solution for European Publication Information Infrastructure.

3 Interoperability and data models

From the ENRESSH-VIRTA-PoC, a set of classes, attributes, and associations was observed that could make for a so-called "lowest common denominator" – a way to unify information from all sorts of source systems and thus achieve metadata that could be compared and analyzed across various countries. The proof of concept also produced a summary of minimum data model elements needed in metadata integration from CRISes and national databases. The minimum data model relies on the CERIF data model which allows representation of various research entities, their connection, and outputs with their semantic relationships. CERIF – the Common European Research Information Format – is a standardized data model for the representation of research entities and their relationships. CERIF is developed by euroCRIS (www.eurocris.org) and is used by various national and organizational CRIS systems.

A subset of metadata from CERIF was chosen (Table 1). The control over which elements should be included in the publication metadata was defined and the elements

were classified as either "Mandatory", "Conditional", or "Optional". For each element, the relevance and context-related issues are made explicit to support the data validation and also add to the bottom-up discussion on source systems on how to make metadata comparable and unified.

Table 1. Proposed metadata for the European Publication Information Infrastructure.

Mandatory	***Conditional****	***Optional***
Publication	ISSN	Audience
Internal identifier	ISBN	DOI
Publication type	Source title	Volume
Publication title	Peer review status	Number
Publication date		Start page
Author		End page
Author's affiliation		
Discipline		

*) Dependent on the type of publication. E.g. a book chapter should be accompanied by the ISBN and the source title of the book.

4 Publication channel databases

According to the ENRESSH-VIRTA-PoC, the core bibliographic data is consistent across countries. However, the extent of other information as well as definitions of open access, publication types, and scientific fields vary. Therefore, a solution based on agreement and shared typologies is not realistic. Instead, the additional information should be determined by using other sources such as the authorized publication channel lists (Figure 3). In many countries, the national publication databases are accompanied by comprehensive authority lists of publication channels, i.e. the journals in and publishers with which the publications are published.

The publication channel databases should also be integrated at the European level. They could be used in determining publication types, peer-review status, open access status, and scientific fields of the publications, and thereby, they help to ensure consistency of national publication databases [21]. Also, other external data sources can be used to check the open access status of publications and machine learning algorithms can be developed to identify the subject fields of the publications. Consequently, there is no need for unifying national field classifications which are often provided under a national framework and sometimes even linked to national legislation.

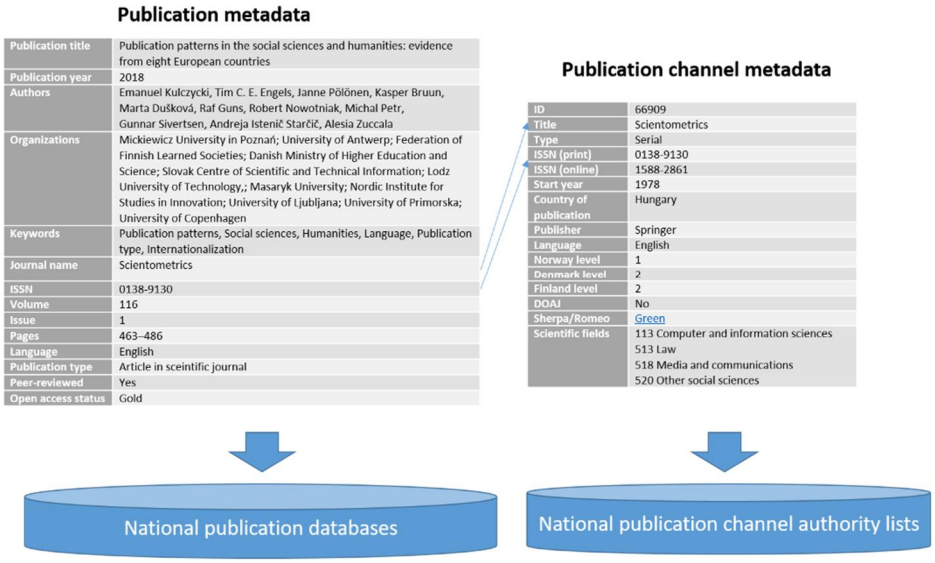

Fig. 3. Example of a publication's metadata, and a publication channel's metadata currently included in the national databases.

5 Conclusions

National publication databases are the most comprehensive data sources of scholarly publications at the national level. In cross-country comparisons, they are, however, underused. Their integration would provide a comprehensive and well-structured information base of European scholarly publications and a great potential for research evaluation at the European level. National publication databases complement other information sources, such as OpenAIRE and Crossref, to provide a complete picture of European research.

However, common standardization of data needs to be defined to have real comparability between research outputs reported to institutional, national, or international databases. Since the core bibliographic information is quite consistent across the databases the data can be enriched by using external sources. Especially high-quality information on publication channels is important in achieving more comparable publication data.

References

1. European Commission: *Assessing Europe's University-Based Research: Expert Group on Assessment of University-Based Research. Directorate-General for Research*. (2010) https://ec.europa.eu/research/science-society/document_library/pdf_06/assessing-europe-university-based-research_en.pdf

2. European Parliamentary Research Service. *Measuring scientific performance for improved policymaking. Science and Technology Options Assessment*. PE 527.383. (2014) http://www.europarl.europa.eu/RegData/etudes/etudes/join/2014/527383/IPOL-JOIN_ET(2014)527383(SUM01)_EN.pdf

3. Waltman, L. *Open Metadata of Scholarly Publications. Open Science Monitor Case Study*. (European Commission, July 2019) https://ec.europa.eu/info/sites/info/files/research_and_innovation/open_metadata_of_scholarly_publications_0.pdf

4. Martin, B., Tang, P., Morgan, M., Glänzel, W., Hornbostel, S., Lauer, G., … Žic-Fuchs, M. *Towards a Bibliometric Database for the Social Sciences and Humanities - A European Scoping Project*. (2010) https://globalhighered.files.wordpress.com/2010/07/esf_report_final_100309.pdf

5. European Commission. *COMMISSION RECOMMENDATION (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information.* (2018) https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018H0790&from=EN

6. Science Europe. *Plan S: Accelerating the transition to full and immediate Open Access to scientific publication*s. (2018) https://www.coalition-s.org/

7. Sivertsen, G. *Scientometrics*, **107**:2, 357–368 (2016) https://doi.org/10.1007/s11192-016-1845-1

8. Aksnes, D. W. & Sivertsen, G. (2019). *Journal of Data and Information Science*, **4**:1, 1–21. https://doi.org/10.2478/jdis-2019-0001

9. Gusenbauer, M. *Scientometrics* **118**, 177–214 (2019). https://doi.org/10.1007/s11192-018-2958-5

10. Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., ... & Haustein, S. PeerJ, **6**, e4375. https://doi.org/10.7717/peerj.4375

11. Gorraiz, J., Melero-Fuentes, D., Gumpenberger, C., & Valderrama-Zurián, J.-C.. *Journal of Informetrics*, **10** :1, 98–109 (2016) https://doi.org/10.1016/j.joi.2015.11.008

12. Boudry, C., & Chartron, G. *Scientometrics*, **110**:3, 1453–1469 (2017). https://doi.org/10.1007/s11192-016-2225-6

13. Fasae, J. K. & Oriogu, C. D. *Library Philosophy and Practice*. **1785** (2018) https://digitalcommons.unl.edu/libphilprac/1785

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

15. *The Declaration on Research Assessment (DORA)*. (2012). https://sfdora.org

16. Hicks, D., Wouters, P. F., Waltman, L., de Rijcke, S., and Rafols, I. *Nature*, **520**:7548, 429–431 (2015). https://doi.org/10.1038/520429a

17. Wilsdon, J. et al. *The Metric Tide. Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. HEFCE (2015). Https://doi.org/10.13140/RG.2.1.4929.1363

18. Kulczycki, E., Engels, T., Pölönen, J., Bruun, K., Duskova, M., Guns, R., Nowotniak, R., Petr, M., Sivertsen, G., Starčič, A., & Zuccala, A. *Scientometrics*, **116**:1, 463-486 (2018). https://doi.org/10.1007/s11192-018-2711-0

19. Kulczycki, E., Guns, R., Pölönen J., Engels, T., Rozkosz, E. & Zuccala, A., Bruun, K., Eskola, O., Starčič, A.I., Petr, M. & Sivertsen, G. *Journal of the Association for Information Science and Technology,* (2020). https://doi.org/10.1002/asi.24336

20. Puuska, H. M., Guns, R., Pölönen, J., Sivertsen, G., Mañana-Rodríguez, J., & Engels, T. *Proof of concept of a European database for social sciences and humanities publications: description of the VIRTA-ENRESSH pilot*. ENRESSH report (2018). https://doi.org/10.6084/m9.figshare.5993506

21. Sivertsen, G. Developing Current Research Information Systems (CRIS) as data sources for studies of research. in Glänzel, W., Moed, H.F., Schmoch, U., Thelwall, M. (Eds.), *Springer Handbook of Science and Technology Indicators*. Cham: Springer, 667-683 (2019). ISBN 978-3-030-02511-3.