Proof of concept of a European database for social sciences and humanities publications:
Description of the VIRTA-ENRESSH pilot

Report

March 2018

Proof of concept of a European database for social sciences and humanities publications: Description of the VIRTA-ENRESSH pilot

Hanna-Mari Puuska[a], Raf Guns[b], Janne Pölönen[c], Gunnar Sivertsen[d],
Jorge Mañana-Rodríguez[e], Tim Engels[b]

[a]CSC – IT Center for Science, Keilaranta 14, Espoo (Finland), hanna-mari.puuska@csc.fi

[b] Centre for R&D Monitoring, Faculty of Social Sciences, University of Antwerp, Middelheimlaan 1, Antwerp, 2020 (Belgium), raf.guns@uantwerpen.be, tim.engels@uantwerpen.be

[c] Federation of Finnish Learned Societies, Snellmaninkatu 13, Helsinki, 00170 (Finland), janne.polonen@tsv.fi

[d] Nordic Institute for Studies in Innovation, Research and Education, P.O. Box 2815, Oslo, 0608 Toyen (Norway), gunnar.sivertsen@nifu.no

[e]Philosophy Institute, Spanish National Research Council, C/Albasanz, 26-28, Madrid 28037 (Spain), jorge.mannana@cchs.csic.es

doi:10.6084/m9.figshare.5993506
Helsinki: CSC & ENRESSH.

Acknowledgment

This report is the result of collaborative work within the COST Action ENRESSH (CA1537 "European Network for Research Evaluation in the Social Sciences and the Humanities"). The authors thank the VIRTA steering group in the Finnish Ministry of Education and Culture for its support of this proof of concept.

Suggested reference

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

Contents

1 Introduction 5

2 The VIRTA system 6

2.1 Origins and use in Finland 6

2.2 Features 7

2.3 Potential for use at the European level 8

3 Implementation of the VIRTA-ENRESSH-POC 9

3.1 Participating institutions 10

3.2 Data format and requirements 10

3.3 Publication types 10

4 Decisions/steps taken at the level of each partner 12

4.1 Three Finnish universities 12

4.2 University of Antwerp (Flanders, Belgium) 12

4.3 University of Oslo (Norway) 13

4.4 University Carlos III Madrid (Spain) 13

5 Description of the data 14

6 Potential improvements 15

7 Conclusion and potential as a science policy supporting tool 17

8 References 19

Appendix A. Data field requirements in the pilot 20

1 Introduction

The European Network on Research evaluation in Social Sciences and Humanities is a EU-funded COST-network with partners from 36 European Countries (www.enressh.eu). The network aims at advancing the understanding of SSH research through three working groups: (1) a working group regarding conceptual frameworks for SSH research evaluation, (2) a working group regarding the societal impact of SSH research, and (3) a working group on databases and the uses of data for understanding SSH research. We here report on a collaborative project conducted as part of the activities of the working group 3 involving partners from Belgium, Finland, Norway, and Spain.

The idea for the project stems from the constant finding in bibliometric research that the most widely used commercial databases, Web of Science and Scopus, do not provide complete coverage of research output in any field, and that in SSH fields they suffer from severe lack of coverage of publications in books and languages other than English (e.g. Sivertsen 2016).

Science policy and research evaluation at all levels of the European Research Area need support from reliable, comparable, and comprehensive information on research activity, productivity and quality. To this end, an Expert Group on Assessment of University-Based Research recommended in a report to the European Commission that it should "Invest in developing a shared information infrastructure for relevant data to be collected, maintained, analyzed, and disseminated across the European Union" (European Parliament 2010).

According to a report to the European Parliamentary Research Service (Mahieu, Arnold & Kolarz, 2014), 19 European Union Member States had developed or were developing a national research information system. The report recommends development of a European integrated research information system inter-connecting the existing national research information systems. To facilitate such integration Science Europe (2016) invites all research organizations to develop resilient information systems in line with the principles of flexibility, openness, FAIRness, and data entry minimization. An important source of experience and learning in this respect, covering possible obstacles and achievements as well as necessary considerations in the process, is the national Research Core Dataset project for the science system in Germany (Biesenbender and Hornbostel, 2016).

A database for the social sciences and humanities (SSH) outcomes is a crucial component of the European research information infrastructure. In a report to the European Science Foundation and the British, Dutch, French, and German research councils, a European Scoping Project recommended, in order to achieve complete coverage of the SSH scholarly output, either negotiations to expand and/or create a new database with the suppliers of Web of Science, Scopus and Google Scholar, or integration of data from national and institutional research documentation systems (Martin et. al. 2010).

The main difficulty of standardization and interoperability of data at the European level is the variety of institutional and national publication information systems and their data models. Many countries are facing a similar problem at national level when they compile information from research organizations using various local systems. Therefore, the Memorandum of Understanding of the COST Action ENRESSH (http://enressh.eu/wp-content/uploads/2016/10/CA15137-e.pdf) stresses the need for coordination at the European level. More specifically the MoU sets forward as tasks for the ENRESSH Working Group 3 'Databases and uses of data for understanding SSH research' the "Development of common rules and procedures for building databases [of social sciences and humanities publications]", and the "Design of a roadmap for a European bibliometric database" (MoU, p 14). As a step towards the latter task, we set up a proof of concept of a European database for social sciences and humanities publications. This process, which builds on the strengths of the Finnish VIRTA system, is described in this report. We refer to this proof of concept as the VIRTA-ENRESSH-POC.

The report is structured as follows. First we introduce the VIRTA-system and its potential for use at the European level. Then we present the setup of the VIRTA-ENRESSH-pilot and the steps taken by the six participating institutions representing four countries. In the final two sections we provide an overview of the data collected and the potential improvements. We conclude with a discussion of the future potential, including the potential as a science policy tool.

2 The VIRTA system

2.1 Origins and use in Finland

The Finnish Ministry of Education and Culture compiles bibliographic information annually of all scientific publications from Finnish higher education institutions and other research organizations. The publication data collection commenced in 2011 and since, each university reports metadata of all its publications to the ministry once a year. The national data are used in monitoring research and it supports the performance-based research funding system (PRFS) for allocation of block-grant funding to universities.

National data collection covers 54 organizations, including 14 universities, with almost 60,000 scholarly publications per year. The publication data include also outputs specific to SSH, such as national language literature and publications aimed at non-scholarly audience.

An advanced decentralized solution to integrate institutional data at the national level, the *VIRTA Publication Information Service*, was launched in the spring 2016. In VIRTA, the Finnish organizations store a copy of publication information of their institutional CRISes or other publication databases. The organizations use various local solutions for publication data collection, such as commercial Current Research Information Systems (CRIS), self-made publication registers, institutional publication repositories and e-forms.

HEIs and other research organizations are responsible for data collection to their local CRISes and have various data collection practices. Import from international and national publication databases, researchers themselves, as well as from library and data-collection personnel can be involved in registering and validating the data contents to the local CRISes.

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.

The publication metadata can be exported to other services or systems via REST API or OAI-PMH from VIRTA, and all data are openly available through JUULI portal with links to full texts (www.juuli.fi). Statistical data are available in www.vipunen.fi .

The publication metadata can be used by research funders, publication or data repositories, infrastructure services, or any other service used by researchers. The major Finnish research funding organization, Academy of Finland, introduced import from VIRTA in its project reporting service in the spring 2017.

2.2 Features

Data are transferred from local CRISes to VIRTA as XML files at least once a year, but many organizations have set a daily automatic update from their own system. All data from previous years to present can be transferred. The organizations transfer their data into VIRTA via a secure and certified connection by using SFTP protocol and SSH authentication keys.

If not able to produce an XML from its own system, the organization can also use the VIRTA CSV-XML tool in order to convert a CSV file into XML. For such organizations that do not have their own CRIS or other publication information system, CSC has developed the JUSTUS Publication Data Input Service, which was implemented in 2017.

The content of the XML format is defined in the VIRTA XML Schema (available online at https://confluence.csc.fi/display/VIR/XML-Skeemat).

To be admitted from the local CRISes to the VIRTA publication information service, HEIs need to provide for each publication type certain obligatory or voluntary data contents (see Table 1), and the data needs to fulfil certain technical criteria (correct form of ISSN, ISBN, etc). All the definitions and requirements concerning the data contents have been stated by the Ministry of Education and Culture in the *Publication data collection Guide* to ensure uniformity and quality of data.

Table 1. Metadata contents in VIRTA. Mandatory and optional fields.

Mandatory in data transfer	Optional in data transfer but mandatory in Ministry's data collection in Finland	Optional in Ministry's data collection	Other information not included in Ministry's data collection	Information generated automatically in VIRTA
Organization ID Publication ID Publication year Title Authors Publication type Authors in organization	Number of authors Scientific field (1-6) International (yes/no) International co-publication (yes/no) Co-publication with a company Open Access ISBN* ISSN* Conference title* Publisher*	Volume Number Pages Article number Journal/series title Country Unit or department Place of publishing Host publication's title Host publication's editors Language DOI Permanent address Source ID Keywords ORCID	Project ID Funder ID	Reporting year Publication's VIRTA ID Co-publication's VIRTA-ID Publication Forum ID Publication Forum ranking Status of the publication

*) Requirements depend on publication type.

Missing fields, incorrect data as well as inter-organizational co-publications and duplicates are automatically identified in VIRTA and the organizations receive instant checking reports. The identification of duplicates and co-publications is based on the following fields:

Priority of the rule	Rule	Publication types
1	DOI	All
2	ISSN + volume + number + pages + publication title	All
3	publication type + publication title + publisher	Monographs
4	Host publication's title + publication title	Book chapters
5	ISBN + publication title	All

Duplicates, errors as well as inter-organizational co-publications are identified automatically and in real time. Error reports are available for research organizations in an online service.

Also the publication channel, if included in the Finnish Publication Forum authority list of journals/series, conferences and book publishers, is identified automatically on basis of ISSN and ISBN codes as well as the conference or book publisher names. The publications are also given a Publication Forum channel ID and level rating (www.julkaisufoorumi.fi).

2.3 Potential for use at the European level

The extension of the Finnish VIRTA publication information service to other European countries and institutions is a potential solution for a European decentralized system aimed at integration and visibility of data about and for the SSH and other fields of science. A copy and modification of the Finnish technical solution and contextual definitions would be a convenient and cost-efficient way of developing a European system.

The envisioned European Research Information Service built upon the VIRTA concept would provide a complete overview on European research publications including all types of scholarly publications and potentially other research outcomes in the future as well. As any country or institution would be able to join, the barrier to adoption would be low, in turn potentially leading to good coverage at the European level.

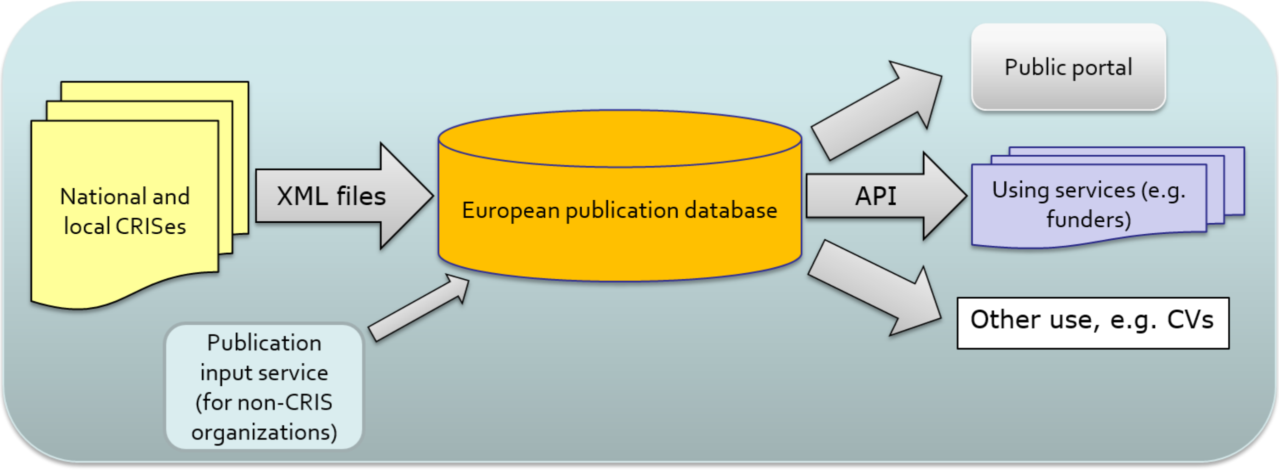
National and local CRISes
XML files
European publication database
Publication input service (for non-CRIS organizations)
API
Public portal
Using services (e.g. funders)
Other use, e.g. CVs

Figure 1. Schematic overview of envisioned European Research Information Service

The technical checking procedure along with well-defined shared standards would ensure the commensurability and uniformity of the data. However, a simplification and modification of the data content definitions that are currently used in VIRTA would be needed, as some of them are specific to the data collection needs in Finland. The quality of data is an ongoing process which will improve over time as the participating organizations and countries would adjust for the shared system little by little. The quality is also supported by the transparency of data when the institutions see, check and compare their and other institutions' data.

Figure 1 provides a schematic overview of how the system would work. At the center is the VIRTA-based system with a database containing metadata on publications from different European countries. National and institutional repositories and CRIS systems can add data to the database by exporting to a standardized XML format. The database can be consulted through a public portal. In addition, the data would be usable in various processes such as researchers' CV's, funding applications, research evaluations, research administration, science policy planning and decision making, research and information retrieval. Through the open APIs, the data can be used in digital services, for example, in importing publications to funding organizations reporting systems.

The service can also be connected to other European systems such as OpenAIRE (www.openaire.eu*)*. The European Research Information Service could either provide metadata to OpenAIRE or receive data from OpenAIRE in order to supplement the national or institutional data. One possible route towards interoperability with OpenAIRE would be to implement the Common European Research Information Format (CERIF) data model (see section 6), since OpenAIRE can import the CERIF XML exchange format (Houssos, Joerg, & Dvořák, 2015).

A VIRTA-based European research information service can also contribute to the Open Science agenda (OECD, 2015) by making publication metadata from European countries available as Open Data. The VIRTA architecture enables publishing the metadata in different formats, including as Linked Open Data, XML, and CSV. Open Data could be published at the level of individual publications if the data providers agree to the publication of their metadata under a permissive license like Creative Commons Zero (CC0, https://creativecommons.org/publicdomain/zero/1.0/) or Open Data Commons (https://opendatacommons.org/licenses/). Even if a more restrictive license applies to (some of) the source data, it would still be possible to publish aggregated statistics as Open Data, which in turn could be used in other projects.

3 Implementation of the VIRTA-ENRESSH-POC

The ENRESSH working group 3 *Databases and uses of data for understanding SSH research* together with CSC – IT Center for Science[1] launched the VIRTA-ENRESSH proof of concept project during the ENRESSH meeting in March 2017 in Sofia, Bulgaria. The target of the pilot was to apply the concept to institutional data from various countries and illustrate the potential of the concept for a European Research Information Service.

In December 2016 the proposal was presented in the VIRTA steering group in the Finnish Ministry of Education and Culture, which considered that VIRTA concept could potentially be extended to European level and the technical solution implemented by CSC. The Ministry allocated one person month of work from CSC to the European VIRTA pilot project.

The proposal was also presented to the EuroCRIS steering group for feedback and possible collaboration in view of the OpenAire call for services. In view of achieving the aim of a European

[1] CSC – IT Center for Science Ltd. is a Finnish center of expertise in ICT that provides services for research, education, culture, public administration and enterprises. CSC is a non-profit organization owned by the Finnish state and higher education institutions.

system the initiators of this report will continue working with EuroCRIS, OpenAIRE and any other initiatives that aim at integration of publication metadata, in particular those pertaining to the social sciences and humanities, across European countries.

3.1 Participating institutions

Six universities from four European countries participated in the VIRTA-ENRESSH-pilot. These universities are:

- University of Helsinki, Finland
- University of Jyväskylä, Finland
- Tampere University of Technology, Finland
- University of Antwerp, Flanders, Belgium
- University of Oslo, Norway
- University Carlos III Madrid (UC3M), Spain

Publication metadata for the years 2014 and 2015 were included in the pilot. Each university submitted either all its publication metadata or the publications metadata from the social sciences and humanities only. Each institution classified all submitted publications into disciplines according to the OECD Fields of Science classification (OECD, 2007).

3.2 Data format and requirements

XML was chosen as the data exchange format, with the Finnish VIRTA XML format as the starting point. To facilitate the implementation for the non-Finnish partners, CSC created a simple CSV model for the pilot, which the participating pilot universities could use to structure their own data. In addition, CSC built a tool that converted these CSV files to the right XML format.

In the pilot, the XML datasets were sent to CSC by email. CSC uploaded the files into VIRTA.

Based on a comparison of the data contents in each participating university and country, the "lowest common denominator", that is, the data fields that all participating universities and countries could supply, was identified. These data fields were set as mandatory for all publications (see descriptions in Appendix A):

- Organization ID
- Organization-specific ID of publication
- Publication year
- Publication title
- Publication authors
- Publication type
- Field of science of the publication
- Organization authors

In addition, there were 28 optional fields, such as ISBN, journal name, open access status, and ORCID. Where possible, these were filled in because they are often crucial in helping to determine whether two records actually refer to the same publication. Finally, there are 3 fields for which a value is automatically generated by the VIRTA system.

3.3 Publication types

So far, there is no shared international standard for publication types. The publication type classifications differ also between the piloting institutions. Analogous categories can however be found in all countries, as most publication databases in Europe use publication type structure including journal articles, books/monographs, edited volumes/anthologies, articles in books and articles in conference proceedings (Sīle et al. 2017).

The current validations and identification algorithms in VIRTA are heavily dependent on the publication types. Therefore, in order to require the least effort, the Finnish publication type classification was chosen as the basis for the publication type classification. UC3M was able to convert their own publication types to the Finnish classification, although not all the Finnish types are used in Madrid. The other countries reported their data according to their own classifications and they were mapped into Finnish classification in VIRTA.

The following publication type mapping was applied:

Finnish universities	UC3M	University of Antwerp	University of Oslo
A1 Journal article (refereed), original research	A1 Journal article (refereed), original research	VABB-1: journal article, peer-reviewed	3= Article in series (ISSN)
A2 Review article, Literature review, Systematic review			
A3 Book section, Chapters in research books	A3 Book section, Chapters in research books	VABB-4: book chapter, peer-reviewed	2= Article in book (no ISSN)
A4 Conference proceedings	A4 Conference proceedings	VABB-5: proceedings paper, peer-reviewed	
B1 Non-refereed journal articles		VABB-1: journal article, non-peer-reviewed	
B2 Book section		VABB-4: book chapter, non-peer-reviewed	
B3 Non-refereed conference proceedings		VABB-5: proceedings paper, non-peer-reviewed	
C1 Book	C1 Book	VABB-2: monograph, peer-reviewed	1= Monograph
C2 Edited book, conference proceedings or special issue of a journal		VABB-3: edited book, peer-reviewed	
D1 Article in a trade journal			
D2 Article in a professional book (incl. an introduction by the editor)			
D3 Professional conference proceedings			
D4 Published development or research report or study	D4 Published development or research report or study		
D5 Textbook, professional manual or guide			
D6 Edited professional book			

E1 Popularised article, newspaper article			
E2 Popularised monograph		VABB-2: monograph, non-peer-reviewed	
E3 Edited popular book		VABB-3: edited book, non-peer-reviewed	
G4 Doctoral dissertation (monograph)			
G5 Doctoral dissertation (articles)			

4 Decisions/steps taken at the level of each partner

4.1 Three Finnish universities

For each of the three Finnish universities involved in this pilot, the VIRTA infrastructure was the basis for data collection. CSC first launched a call for Finnish pilot organizations. Permissions for data use from the three pilot organizations (Helsinki, Tampere Tech, Jyväskylä) was obtained. Then the following steps were taken:

- Create a new database created for the pilot data
- Copy the data from the three Finnish organizations into pilot database
- Translate VIRTA fields into English
- Create a revised CSV-XML converter
- Create revised checking algorithms for mandatory fields in the pilot
- Create revised duplicate identification algorithm
- Validate of the data from the other countries
- Download the data from the other countries
- Run the checking and duplicate identification programs

4.2 University of Antwerp (Flanders, Belgium)

For the non-Finnish universities a detailed description is provided for University of Antwerp. A more high level description is provided for University of Oslo and University Carlos III Madrid. For University of Antwerp the steps followed are as follows:

- The VABB-SHW, a Flemish database of publications from the social sciences and humanities, was chosen over the institutional repository, since it seems likely that a future European Research Information Service would receive data from the regional database rather than from each institutional database separately.
- Set Status of publication to 1 if the publication is included in the WoS or GP selection (i.e., if publication ‘counts’ for the PRFS) and to 0 otherwise
- Restrict set to publications from 2014–2015 and from University of Antwerp
- Add author or editor names, separated by semicolons. Reformat author/editor names to look like, e.g., ‘Van Petegem, Peter’ instead of ‘Petegem, Van, Peter’. A few publications have no author/editor but a different role like ‘collaborator’; add those as authors.
- Add number of authors (the number of semicolons + 1 in the list of authors) and organization authors

- Add pages, formatted like this '*[start page]-[end page]*'. If either is missing, just fill in the page number that is available. Otherwise, leave blank. If article number is known, add that to the Article number field.
- For publication language, map the ISO 639-2 abbreviations used in VABB-SHW to the ones used in VIRTA, e.g. 'dut' to 'nl'.
- For parent publication's title, use the field btitle (book title) and if that is not available, ptitle (proceedings title). For publication title, use field ctitle. For journal name, use field jtitle.
- Abbreviate the Flemish publication types to V1, V2 etc.
- Add ISSNs and/or ISBNs where applicable. For this, use the first ISBN and at most the four first ISSNs.
- Add publisher name, using the ISBN prefix of each given ISBN.
- Add WoS identifiers to Source database code.
- Add FOS fields in numeric format (e.g. '6.5'), following the process described by Guns et al. (2018).
- Finally, add bibliographic details: publication year, volume, issue, place of publishing, publisher, article number, DOI, handle (permanent address).

4.3 University of Oslo (Norway)

Complete data on peer-reviewed scholarly publications were extracted from the Norwegian Science Index (part of Cristin database), limiting to University of Oslo in the publishing years 2014-15.

We were able to provide standardized information for all publications for the following VIRTA variables:

- Organization ID
- Reporting year
- Publication type
- I Field of science of the publication (based on FOS, OECD)
- Organization authors
- Publication authors
- Publication title
- Journal name
- ISSN
- Publisher
- Organization-specific ID of publication

4.4 University Carlos III Madrid (Spain)

- Receive publication data from UC3M (Getafe Campus) internal publication register (This is done by personnel at UC3M).
- Set the common fields for the five publication types available in a single spreadsheet
- Add the specific fields in independent columns of the spreadsheet
- Review cases with one or more compulsory fields missing
- Check ISSN format (length and structure)
- Check ISBN format (length and structure)
- Calculate pages from starting-ending pages columns. Blank if at least one empty.
- Check authors separator (semicolon)

- For field of science, identify single cases, correct potential spelling problems for automatic replacement, develop a correspondence with UNESCO codes and replace names with codes.
- Check field of science codes with master list
- Add 'Unknown' to compulsory fields with no value (institution code, i.e.)
- Search for undue ASCII characters and remove (often found in the pages fields)
- Use the validation option of the converter tool. If error can be corrected: correct. If error cannot be corrected: null.

5 Description of the data

The database that was compiled for the pilot consists of bibliographic metadata of 52,948 metadata records on scholarly publications of 6 research institutions for the period of 2014-2015. Table 2 provides an overview of the number of publications per institution per year. The total number of publications for all six universities is, in fact, lower than the numbers reported here, since co-publications between institutions may be counted more than once. Unfortunately, it was not possible for all institutions to supply sufficiently detailed metadata to make reliable co-publication detection possible. This highlights the importance of good metadata: only when sufficient metadata are available is it possible for the system to adequately detect such duplicate records.

Table 2. Number of publications per pilot institution per publication year

Institution	Publications 2014	Publications 2015
University of Helsinki, Finland	11 987	12 113
University of Jyväskylä, Finland	3 140	3 273
Tampere University of Technology, Finland	1 891	1 921
University of Antwerp, Flanders, Belgium	2 307	1 990
University of Oslo, Norway	5 352	5 491
University Carlos III Madrid (UC3M), Spain	1 800	1 683
Total*	26 477	26 471

* Co-publications between organizations are counted multiple times in the total counts.

For the University of Oslo and the three Finish universities publications from all fields where included. For the University of Antwerp and University Carlos III Madrid only publications from the social sciences and humanities were included. Also, the coverage of non-peer reviewed publications may differ between the research information systems from which the bibliographic data originate. Although in all cases the possibility for inclusion of all major publication types is provided, comprehensiveness of data regarding non-peer reviewed publications depends more on authors reporting these items than is the case for peer reviewed publications.

The distribution of the publications across the publication types is presented in Table 3. As described in section 3.3 a convenience mapping of publications to the publication types distinguished in the VIRTA database was applied. Hence some publication types only occur in the data from Finnish universities and the University Carlos III Madrid.

Table 3. Number of publications per type per publication year

Publication type	Publications 2014	Publications 2015
A1 Journal article (refereed), original research	13 144	13 549
A2 Review article, Literature review, Systematic review	397	441
A3 Book section, Chapters in research books	2 969	2 735
A4 Conference proceedings	2 158	1 825
B1 Non-refereed journal articles	1 735	1 767
B2 Book section	989	921
B3 Non-refereed conference proceedings	315	207

C1 Book	340	323
C2 Edited book, conference proceedings or special issue of a journal	292	263
D1 Article in a trade journal	913	869
D2 Article in a professional book (incl. an introduction by the editor)	196	409
D3 Professional conference proceedings	40	158
D4 Published development or research report or study	483	394
D5 Textbook, professional manual or guide	92	96
D6 Edited professional book	11	66
E1 Popularised article, newspaper article	1 377	1 375
E2 Popularised monograph	169	137
E3 Edited popular book	95	94
G4 Doctoral dissertation (monograph)	326	215
G5 Doctoral dissertation (articles)	436	627
All publication types*	26 477	26 471

* Co-publications between organizations are counted multiple times in the total counts.

In terms of language, bibliographic metadata of publications in any language could be reported. Table 4 provides an overview of the most common languages of the publications included in the pilot database.

Table 4. Publication language of the publications per year

Publication language	2014 (N)	2014 (%)	2015 (N)	2015 (%)
English	15 329	57.90	15 760	59.54
Finnish	4 985	18.83	5 077	19.18
Dutch	1 175	4.44	986	3.72
Norwegian	241	0.91	230	0.87
Spanish	312	1.18	287	1.08
Multiple languages	1 033	3.90	912	3.45
Missing	3 402	12.85	3 219	12.16
Total*	26 477	100.00	26 471	100.00

* Co-publications between organizations are counted multiple times in the total counts.

In sum, for the VIRTA-ENRESSH-pilot all publication types from all disciplines in all languages could be submitted. In view of analysis of the data analysis we identified the analysis of peer reviewed publications, in particular journal articles, as most appropriate and feasible. Such analysis was taken on in the frame of an ENRESSH-short term scientific mission (ENRESSH-STSM) by Joshua Eykens to CSC, Espoo, Finland, and will be reported on in the near future.

6 Potential improvements

Different publication type classifications are used in different source systems. For the participation of institutions in the pilot this was not a hindrance. As the most common publication types (e.g. journal articles, book chapters) occur in almost all systems a common denominator can be found for a substantial share of the publications. For comparability and benchmarking of publication sets, however, a more fine-tuned mapping of publication types may be needed. Moreover, the VIRTA-system uses publication types as a basis for some of its validation and identification processes. A traditional solution, in which agreement amongst all participating institutions and countries is strived for is probably not feasible. Even if such agreement seems feasible among the current six institutions, the need for an agreement might conflict with the national requirements imposed on these institutions (e.g. the publication types described in the current report apply to all Finnish universities). Moreover, any agreement would be temporary as a need for re-negotiation would occur every time that an institution or country would want to add its data.

A related issue pertains to terminology, especially when referring to characteristics of publications. Even if all partners agree on a specific publication type classification, agreement might be needed on the precise meaning and operationalization of terms such as ‘scholarly’, ‘scientific’, and ‘peer reviewed’. As such agreement is currently not a reality (Pölönen, Engels, Guns & Verleysen, 2017) it might be difficult to achieve.

We therefore believe that a more structural approach will be needed. On the technical level, we envision that the validation and identification algorithms in the VIRTA system could be adapted so as to work with less structured data, without losing in terms of quality control. Upon future expansion of the database, this need will be taken into account. On a more structural level, we advocate an ontology-based data management approach (Daraio et al, 2016) to be implemented in the central VIRTA system and across source databases, hence facilitating future expansion of the dataset. As by far most future partners, like the current partners, will start from existing research information systems, the implementation of an ontology-based approach may not be feasible upfront but might gradually be implemented among participating countries and institutions, thus greatly facilitating the integration of data at European and even international level.

The ontological approach also supports making data exchangeable with current research information standards such as EuroCRIS’s CERIF data model. In an ontology-based approach, an important decision is of course the choice of ontology. Here, various factors are relevant, such as expressiveness, domain-specificity, broadness, and adoption elsewhere. The CERIF interchange format, maintained by EuroCRIS, is a logical candidate, given its adoption in various European (CRIS) systems, high level of sophistication, and broad coverage of research information. The price is that CERIF is fairly complex and hence nontrivial to implement, which may be problematic for some, especially smaller, institutions. Instead of forcing data sources to provide their data according to a unique data model another option is to modify VIRTA into a more integration based solution. That is, VIRTA could act as a “data broker” which could map and parse data from various sources and different formats, which would be a much easier solution for data providers.

In terms of data format, further streamlining may facilitate the integration of data into the VIRTA system. The format of a DOI for example is defined internationally yet implemented differently in different systems. Although a uniform implementation across systems is obviously advisable, this is not always what happens in practice. Likewise, the reporting of all available metadata is called for, e.g. including abstracts as they may facilitate processes such as identification of duplicates, and enriching of other objects.

The attribution of the publications to OECD-FOS disciplines occurred in the source systems in each of the institutions (Madrid) or countries (Finland, Flanders, Norway). The underlying subject definitions are available in the documentation by OECD which was used as a basis for classification in each of the sources. We recognize, however, that full documentation of the underlying implementation methodologies is not available and may hamper comparability of the data across institutions and countries. For example, refinements of the OECD-FOS-scheme have been implemented in some countries, the classification may be implemented at the publication level or at the publication channel level, and in some systems more than one discipline may be assigned to a publication whereas this is not allowed in other systems. In view of future expansion of the database we will undertake two steps in order to facilitate comparability of the data collected and to facilitate discussion on the harmonization of the implementations of the OECD-FOS-classification system. First, as part of the analysis of the data, the classifications of all publication channels (journals and books) that occur in more than one source will be compared. All incompatibilities in classification will be listed, discussed and resolved amongst the consortium partners. In addition the consortium partners will work on the cross-mapping of different types of classification systems (e.g. Guns et al, 2018) and will explore the possibility of computer facilitated assignment of publications to OECD-FOS disciplines using text analysis techniques. In the future we envision that the attribution of publications to disciplines will be computer facilitated using the metadata of the publications in cases where no local cognitive classification of publications is available. Starting with English language publications and publications for which English language metadata are available, such an approach could facilitate the expansion of the database to other partners, provide additional useful feedback to participants, and contribute to the comparability of the data collected.

The VIRTA-ENRESSH-pilot was set up to integrate bibliographic metadata originating from different research information source systems. Enriching these data with metadata on publication channels, e.g. the classification of journals as peer-reviewed or not, as high-prestige in different national contexts, or with Web of Science and Scopus based impact factors, makes them immediately useful for benchmarking and monitoring at local, regional, national and European level. The presentation of such data sparked immediate interest during an ENRESSH meeting in Finland, 8 November 2017. Therefore, we envision that in the future the work on an integrating research information system may be joined with an integration of information on classification of publication channels (journals, publishers, and conferences). By submitting a simple CSV format to the VIRTA-service, users could then obtain enriched information regarding the set of publication under consideration, e.g. the publications of a whole institution or one of its departments. Moreover, such data could be compared, overall and per discipline, to data from other participating institutions or a selection of them. In sum, we envision an integrated approach in which bibliographic metadata are enriched with metadata regarding the publication channels in which the publications appeared.

7 Conclusion and potential as a science policy supporting tool

A total of over 50 000 references were integrated into the VIRTA-ENRESSH-pilot database. In this report we document the steps taken during this pilot, describe the data collected and discuss potential improvements. The potential as a science policy supporting tool of an integrating service for research information system is obvious: a small concerted investment using existing data from different source systems makes it possible to benchmark and monitor outputs across institutional and national boundaries. The current dataset already allows to contrast publication patterns in terms of volumes, language use and prestige of publication channels between the participating countries and institutions. In the European context with a large linguistic diversity and almost no cross national data integration such possibilities for comparison bear huge potential.

The pilot demonstrates that it is possible to integrate institutional publication data from different countries using the VIRTA model. This required the identification of data fields that all participating institutions and countries could supply (the “lowest common denominator”), In addition to this, the participants could provide optional data (e.g. ISSN and ISBN) that is needed in the identification of duplicates, co-publications and publication channels, and can be used for analyzing and further enriching the data. This pilot used the data model developed for the specific needs of the Finnish PRFS. The next step is to develop a data model specifically for the purpose of integrating institutional or national publication data from different countries. This needs to be done with an eye towards enhancing comprehensiveness, comparability and further use of the data. Although the data model and system should allow inclusion of all relevant scholarly outputs in different fields, it should also have enough metadata and structure to permit relevant subsets of publications to be used in comparisons and benchmarking.

The main challenge is that institutional and national data sources use different data models as well as different data collection and validation procedures. It is possible to harmonize definitions and practices at the regional or national level, especially if data collection is tied to a PRFS. Such harmonization is much more difficult to achieve to in case of data integration between countries. Participants can, for instance, agree on a publication type classification and can map the publication types from institutional and national data sources to its categories. But a given type, such as article in proceedings, may be defined in somewhat different ways in different sources, and a publication's type may be determined using different methods or heuristics. The same holds true for, e.g., field classifications and the defintion and identification of peer-reviewed outputs. Agreed data definitions and classifications of course help to some extent to diminish the substantive variation that is due to different institutional and national models and procedures behind the integrated publication data.

It is also possible to increase the comparability of data by developing automated methods to restructure and reclassify VIRTA data in a uniform way on the basis of the bibliographic metadata as well as information from external sources. An algorithm could be devised to determine publication types in a uniform way for the entire dataset based on identifiers (ISSN, ISBN, DOI), possibly combined with information on the type of serial (journal or monographic series) from the international ISSN Center. Regardless of differences in peer review definitions, existing regional and national authority lists of peer-reviewed journals, series and book publishers can be used and developed to identify peer-reviewed outputs in a uniform way. Despite differences in field classifications used in the sources, OECD fields can be determined in the same way for the publications in VIRTA on basis of journal fields, or using computer-assisted methods. Algorithmic solutions could also be used to determine publication language in a coherent way, whereas data sources may rely on different sources and methods. This way, the system can offer for analysis purposes both the sources' original information as well as uniformized classifications and structures.

8 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.

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*, 108, 441-455, DOI 10.1007/s11192-016-1913-6.

Directorate-General for Research and Innovation (2010). *Assessing Europe's university based research: Expert Group on Assessment of University-Based Research*. European Commission, https://doi.org/10.2777/80193.

Guns, R., Sīle, L., Eykens, J., Verleysen, F.T., & Engels, T.C.E. (2018). A comparison of cognitive and organizational classification of publications in the Social Sciences and Humanities. *Scientometrics*, under review.

Houssos, N., Joerg, B., & Dvořák, J. (2015). OpenAIRE Guidelines for CRIS Managers 1.0 [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.17065

Mahieu, B., Arnold, E., & Kolarz, P. (2014). *Measuring scientific performance for improved policy making. Summary of a study* (p. 16). European Parliament; Directorate-General for Parliamentary Research Services.

Martin, B., Tang, P., Morgan, M., Glänzel, W., Hornbostel, S., Lauer, G., … Žic-Fuchs, M. (2010). *Towards a Bibliometric Database for the Social Sciences and Humanities - A European Scoping Project* (p. 55).

Organisation for Economic Co-operation and Development. (2007). *Revised field of science and technology (FOS) classification in the Frascati manual*. Paris: Working Party of National Experts on Science and Technology Indicators, Organisation for Economic Co-operation and Development.

Organisation for Economic Co-operation and Development. (2015). *Making Open Science a Reality*. OECD science, technology and industry policy papers no. 25. Paris: OECD Publishing. https://doi.org/10.1787/5jrs2f963zs1-en

Pölönen, J., Engels, T.C.E., Guns, R., & Verleysen, F.T. (2017). Is my publication peer reviewed? A comparison of top-down and bottom-up identification of peer review in the framework of the Finnish and Flemish performance-based research funding systems. *STI2017: 22nd International Conference on Science and Technology Indicators* (Paris, 6–8 September 2017).

Science Europe. (2016). Position Statement on Research Information Systems. Co-ordination by the Science Europe Working Group on Research Policy and Programme Evaluation. D/2016/13.324/11. Brussels: Science Europe. http://www.scienceeurope.org/wp-content/uploads/2016/11/SE_PositionStatement_RIS_WEB.pdf

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

Sivertsen, G. (2016). Publication-Based Funding: The Norwegian Model. In M. Ochsner, S. E. Hug, & H.-D. Daniel (Eds.), *Research Assessment in the Humanities* (pp. 79–90). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-29016-4_7

Appendix A. Data field requirements in the pilot

Finnish	English	Definition	More information	Finland, Flan-ders, Madrid	Re-quired field
<OrganisaatioTunnus>	Organization ID	Organization ID according to a certain classification	ID is 5-8 characters and it is determined according to a certain classification. e.g. Finnish universities		X
<IlmoitusVuosi>	Reporting year	The year in which the publication was reported for the first time, 2011 or after.	Reporting year is automatically generated in VIRTA.		
<JulkaisunTilaKoodi>	Status of the publication	Status of the publication. Value -1,0, 1, 2 or 9	Values: -1 -> Rejected 0 -> Accepted with inadequate data 2 -> Accepted with complete data 9 -> co-publication Generated automatically in VIRTA		
<JulkaisunOrgTunnus>	Organization-specific ID of publication	The organization's own ID for the publication.	Free text field, for example sequential numbers, e.g. 2015_0001, 2015_0002, etc.		**X**
<YksikkoKoodi>	Organization sub-unit	Faculties, departments or units (max. 20) of the organisation with contribution to the publication	Free text field. Codes defined by the organization		
<JulkaisuVuosi>	Publication year	The year in which the publication was published for the first time as a version with full bibliographic information.	year between 1900–2020	X	X
<JulkaisunNimi>	Publication title	Publication title as given in the article or the book. If necessary, the title of a foreign-language publication may be transliterated.	Free text field	X	X
<TekijatiedotTeksti>	Publication authors	Authors of the publication in the format and order in which they were listed in the original publication or source database.	Free text field. Several authors should be separated by semicolon: Forename, Surname; Forename, Surname	X	X
<TekijoidenLkm>	Number of authors in publication	The total number of authors in the publication.	A positive integer, not 0.		
<SivunumeroTeksti>	Pages	Publication's page numbers in which the article was published in the same format as in the original article or source database.	Free text field	X	
<Artikkelinumero>	Article number	Article number used for the publication of the article (if applicable) in the same format as in the original article or source database. (Usually in electronic publications)	Free text field		
<AvainsanaTeksti>	Keywords	Keywords that describe the content of the publication as accurately as possible.	Free text field. Several keywords should be separated by semicolon		

Finnish	English	Definition	More information	Finland, Flanders, Madrid	Required field
<ISBN>	ISBN	Publication or parent publication ISBN number.	Validity is verified by the method described in http://isbn-information.com/10-digit-isbn.html and http://isbn-information.com/13-digit-isbn.html	X	
<JufoTunnus>	Publication Forum ID	Publication forum identifier according to the Finnish Publication forum (JUFO-ID) (e.g. 5003).	http://www.tsv.fi/julkaisufoorumi/haku.php		
<JufoLuokkaKoodi>	Publication Forum ranking	The classification has three levels: 1 = basic; 2 = leading; 3 = top (0 = identified publication channels which have not received level 1)	Value generated in VIRTA. http://www.tsv.fi/julkaisufoorumi/materiaalit/jufo_arviointikriteerit.pdf		
<JulkaisumaaKoodi>	Publishing country	Country of publication of the journal, series, monograph or parent publication according to the countries 2007 classification of Statistics Finland.	Value according to the countries 2007 classification of Statistics Finland.		
<LehdenNimi>	Journal name	Journal/series name, as complete as possible, and spelled out (no abbreviations). If the name of a conference article journal/series is unknown, the established conference name will be indicated without the ordinal and year and with no abbreviations.	Free text field	X	
<ISSN>	ISSN	The ISSN number of the series publishing the journal, monograph or parent publication according to the primary printed version. If there is no printed version, the ISSN number of the electronic version will be indicated.	Validity is checked. 1-2 ISSN numbers per publication can be	X	
<VolyymiTeksti>	Volume	Volume of the journal or series in which the article appeared.	Free text field	X	
<LehdenNumeroTeksti>	Issue	Issue of the journal or series in which the article appeared.	Free text field	X	
<KonferenssinNimi>	Conference title	The established name of the conference repeated in the name of the proceedings publication.	Free text field		
<KustantajanNimi>	Publisher	Publisher's name, as complete as possible, and spelled out (no abbreviations).	Free text field	X	
<KustannuspaikkaTeksti>	Place of publishing	The place or places given in connection with the publication's publisher.	Free text field	X	
<EmojulkaisunNimi>	Parent publication's title	Name of the edited book in which the article was published.	Free text field	X	
<EmojulkaisunToimittajatTeksti>	Parent publication's editors	Editors of an edited publication in the format and order in which they were listed in the original publication or source database.	Free text field		
<JulkaisutyyppiKoodi>	Publication type	Publication type according to the publication type classification		X	X

Finnish	English	Definition	More information	Finland, Flan-ders, Madrid	Re-quired field
<TieteenalaKoodi>	Field of science of the publication	One to six fields of science in the order of relevance of each field to the publication. The first, so-called primary field of science is mandatory		X	X
<YhteisjulkaisuKVKytkin>	International co-publication	At least one of the author's is affiliated to a foreign organization	0 = no, 1 = yes		
<JulkaisunKansainvalisyysKytkin>	Internationality of publication* (domestic/international)	The publisher of a national publication is the home country of the reporting organization. The publisher of an international publication is not the home country. The publisher of a conference publication refers to the publishing house.	0 = no, 1 = yes		
<JulkaisunKieliKoodi>	Publication language	The language used to write the publication according to the Languages 2003 classification of Statistics Finland http://www.stat.fi/meta/luokitukset/kieli/001-2003/index.html	https://virkailija.opintopolku.fi/koodisto-service/rest/kieli/koodi	X	
<AvoinSaatavuusKoodi>	Open access	Open access status of the publication	0 = No answer 1 = Publication published on an open access channel (all publications on the channel are openly accessible) 2 = Open access publication published on a hybrid channel (the channel contains both open access and non-open access publications)		
<YhteisjulkaisuYritysKytkin>	Co-publication with a company	At least one of the author's is affiliated to a company	0=no, 1 =yes		
<RinnakkaistallennettuKytkin>	Self-archived	The publication is self-archived in a field-specific or institutional repository	0=no, 1 =yes		
<RinnakkaistallennusOsoiteTeksti>	Self-archived permanent address	Permanent address of a self-archived publication (e.g. URL)	Free text		
<DOI>	DOI	The Digital Object Identifier (DOI) of the publication.	Validity is checked. https://dx.doi.org/	X	
<PysyvaOsoiteTeksti>	Permanent address	Website address based on permanent identifiers (e.g. DOI, URN or handle) of the publication that takes the user directly to the full text version of the publication.	Free text field		
<LahdetietokannanTunnus>	Source database code	Publication identifier or ID number in the database from which its record was harvested (e.g. Web of Science, Scopus, Pubmed, ArXiv, Cab Abstracts, Arto, Fennica).	Free text field	X	
<OrganisaationTekijat>	Organization authors	Author affiliated in the reporting organisation.	Free text field. May include several names but at least one is compulsory. Several authors should be separated by semicolon: Forename, Surname; Forename, Surname		X

Finnish	English	Definition	More information	Finland, Flanders, Madrid	Required field
<ORCID>	ORCID	ORCID identifiers of authors from the reporting organisation, e.g. 0000-0000-0000-0000, see http://www.orcid.org	e.g. 0000-0000-0000-0000, see http://www.orcid.org		