# European Publication Information Infrastructure Data Model

## Introduction

Current developments in institutional and national databases have led to more information about research outputs, especially on publications which have become commonplace output to be used on the research evaluation and funding allocation. Institutional and national databases on publications, be it an institutional current research information systems (later on CRIS) or a national house-built solution, are in almost all cases built for the context of monitoring researcher's outputs in a way that this data can then be used to evaluate on an institutional level (e.g. tenure-track, recruitment) or nationally (e. g. national funding models that take into account the research outputs in form of indicators) and further explored by general public (via e.g. portals or statistical dashboards). What makes both institutional and national databases apart from larger commercial bibliographic databases is that they assertive in including outputs from social sciences and humanities (later on SSH). To some extent also research and research outputs that are highly national (e.g. articles in domestic journals, publications in national language) are not well covered in commercial databases.

Main benefit of CRIS systems in general, be it implemented on an institutional or national level, is the quality of metadata which is achieved by both the highly structured format in which the data on research outputs is collected and the extent in which the metadata is reported from. Well structured metadata enables sophisticated analysis on research and the extent provides for wider perspective - research can be assessed from the smallest sections of research all the way to institutional or even national level.

To achieve a complete and accurate bibliographic database on an European level has been on the agenda for quite a few organizations or projects. European Network for Research Evaluation within the Social Sciences and Humanities (later on ENRESSH) and the Working Group 3. The main objective of ENRESSH Working Group 3 is to reflect upon the standardisation and the interoperability of current research information systems (CRIS) dedicated to research outputs from the social sciences and humanities (SSH). One of the goals is to develop shared procedures for building and maintaining databases and design a roadmap for a European database. As part of the working group's work, a pilot case study was made on utilizing the Finnish national VIRTA Publication Information system's solution for wider set of organizations in Europe. This collaborative VIRTA-ENRESSH-POC of a decentralized approach to aggregate publication metadata was launched in Spring 2016 and the case study was carried out between 2017-2018 for 6 organizations from 4 countries (Belgium, Finland, Norway and Spain). In this POC, it was discussed that if on an European level a so called European Publication Information Infrastructure could be built, that would provide a complete overview, i.e. metadata on publications, and would include all types of scholarly publications from all fields of science. The data collected in the pilot from had its highest quality and consistency in terms of the bibliographic data meanwhile the classifications varied. In this context, the research on publishing in SSH fields can be achieved and wider use of research outputs as a base for analysis on research is made possible.

Perhaps even a bigger undertaking of collecting and combining bibliographic metadata on research publications on European level is ongoing as part the agenda of OpenAIRE, an organization behind a network of open science specialists in Europe and currently hosting one of the largest databases in Europe on research outputs. By utilizing the euroCRIS's Common European Research Information Format (later on CERIF) data model at it's core, OpenAIRE has, with first version dating back to 2015 and recent updates in 2018, gained momentum by the OpenAIRE Guidelines for CRIS Managers to support metadata harvests from various institutional and national publication databases (e.g. VIRTA) and CRIS systems (e.g. METIS,). As it stands, in 2019 several CRIS systems aim to be compliant with the Guidelines and thus harvestable by the OpenAIRE. This compliance is not yet achieved by many, and only a couple of CRIS systems are included in the so-called beta infrastructure of OpenAIRE. From these five systems, VIRTA stands out being the only national CRIS system to provide for OpenAIRE.

Albeit both initiatives seemingly share a common goal of having a complete set of metadata on publications, there is somewhat distinctive difference in the approach on accumulating metadata on research publications. While OpenAIRE already has a high number of records for publications (some 26 mil.), there is a high variation on the coverage of publications e.g. on a national level and a vast majority of publication are harvested either via repositories and other publication aggregators. This, although seemingly a very high number of records, is still far from commercial databases i.e. Scopus, Web of Science and Google Scholar (amount of records ranging from 100 to 400 mil.). OpenAIRE Explore (a portal for exploring the individual publication and its metadata from OpenAIRE database) aims to provide researchers and other interested persons a way to find relevant research. Other services include are the so called Content Provider Dashboards, which make it possible to measure and monitor the contents of your harvested database and make it possible to enrich the harvested metadata even further. OpenAIRE can be seen highly as a top-down implemented system as there is almost no practicable way to have metadata harvested in a way, that could also pay attention to the institutional or national context in which the source system is either built up or working. This extends to many dimensions of the metadata, how organizations are handled, which disciplines are included in the metadata, what kind of publication types are used.

For purposes of exploring research publications, OpenAIRE provides great starting point. However, if there is a need to evaluate, monitor or assess some part of research done on an institutional, national or international level, the contents fall short. Mainly this is related to how the publications have been accumulated to the OpenAIRE database. Each harvested systems has to follow a set of Guidelines provided by OpenAIRE (currently including 3, for Literature Repositories, Data Archives and CRIS Managers) which state the use of data models (for repositories Dublin Core; data archives Data Cite and CRISs CERIF) and what validation there is in place for each metadata value provided in the harvest (format, ranges etc.). Especially for the CRIS Managers Guidelines there is little control on the metadata quality itself as many elements are included as optional and only a bare minimum of metadata is mandatory for the harvest. This leads to cases for example for research publications in which the metadata is minimal, only containing information e.g. on the title and the general publication type of this certain record. On many occasions the source systems, e.g. CRIS systems, have much higher quality metadata available, but via OpenAIRE harvest much of this information is lost due to shortcuts on mapping of the data models, the small amount of resources invested in providing metadata via endpoint to be harvested or the incompatibilities between data models. There is also little guarantee that the metadata on publications is evenly spread among scientific fields or between national and international (i.e. English) language, as some disciplines and publications written in English have tendency to skew the contains of large bibliographic databases.

One major findings of the ENRESSHs projects is that for a research publication database it is of great importance to be able to have a complete and inclusive set of research outputs, be it on institutional or nation level, for it to be used in any form of assessment or evaluation of research. This is generally achieved by the use of context relevant system choices, data models and criteria to import or input the publication metadata to databases. This approach is quite different of those of commercial databases or e.g. OpenAIRE, where metadata requirements are not able to take into account context related documentation on metadata, e.g. criteria on what is determined as "scientific" or what counts as an "article". For this reason, the aggregating commercial databases are in many cases not well equipped to answer to questions like "How many publications does organization X produce?" or "Which scientific field is most prominent in country Y?". Thus, their use in institutional or national contexts is difficult, as the coverage or the quality of metadata do not meet the needs that are set by various research and analysis use cases.

## Data model for European Publication Information Infrastructure

Following the ENRESSH-VIRTA-POC, an idea of required metadata model to be used on European level was discussed. This common standardization and data content would need to be defined to have real comparability between research outputs reported to institutional, national or even international databases. From the POC of 6 organizations and 4 countries, a certain set of classes, attributes and associations were observed that could make for a so-called "lowest common denominator" - a way to unify metadata from all sorts of source systems and thus achieve metadata that could be compared and analyzed across data from various countries in Europe. Thus 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 infrastructure 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.

As one deliverable of this STSM is the further analysis and draft of a data model for European Publication Information Infrastructure. The data model is to be as interoperable as possible, yet aiming to have as high quality metadata as possible.

Interoperability is of crucial importance when the source systems collecting metadata from research are numerous and vary heavily from institution and nation to another. 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 data model, maintained by EuroCRIS, is a logical candidate, given its high level of sophistication, and broad coverage of research information. As another benefit of using CERIF is that there are many systems already aiming for compliance of their system in CERIF format e.g. for use in OpenAIRE harvesting. Thus, the planned data model would be based on CERIF classes and attributes.

High quality metadata would be achieved by extending the mandatory and conditional attributes that are required from the source systems. This approach is similar to what is done on national level in e.g. VIRTA and was further explored in ENRESSH-VIRTA-POC. By having a unified set of attributes and controlled ranges, the usability of data on research outputs is greatly enhanced.

The scope of this data model is currently limited to research publications only. Thus emerging outputs like data sets, software etc. are left out and focus is on traditional publications as it still is arguably the main output of research. For other outputs of research, the standards on e.g. publishing and metadata format are still under change and this would lead to worse quality metadata when aggregated on a broader level.

## Deliverable

1. A summary of minimum CERIF data model elements needed in research publication metadata transfers considering CRIS systems and national aggregators in European context
    a. Summary of European Publication Information Infrastructure Data Model

# Implementing in European Publication Information Infrastructure

In VIRTA-ENRESSH-POC, the concept of Finnish Publication Information Service VIRTA was used as a basis for the technical implementation i.e. a database and metadata exchange was done using the same procedures as in the national system of VIRTA. This system was set up to integrate bibliographic metadata originating from different research information source systems. For POC the VIRTA technical solutions in VIRTA were enough, but for European level system a dedicated infrastructure should take place. This does not prevent using the POC procedures as well - there is certainly a need for e.g. simple CSV-XML tool for publication registries with limited technical possibilities.

Second deliverable of this STSM is to provide an outline of the implementation of research metadata exchanges in European Publication Information Infrastructure. As well as the data model, this is to support the high quality of metadata on research outputs in Europe, but in addition is to be a bottom-up based system, where the context and coverage of source systems are valued highly when aggregating metadata on research outputs.

For high quality metadata, the European Publication Information Infrastructure should, as was discussed in VIRTA-ENRESSH-POC, be expanded on the possibilities of metadata exchanges. Thus in addition to the CSV-XML tool and the drafted separate publication input service, the infrastructure should support the metadata exchange via APIs, more specifically the OAI-PMH protocol currently in use or being adopted by various CRIS systems in Europe. This allows for wider coverage of source systems to be aggregated and covers systems from simple publication registries to full CRIS systems to be included in the European database.

For bottom-up based approach a great emphasis on the source system context and validation on the aggregation should be in place for the European Publication Information Infrastructure, as this is important for the data comparability and uniform metadata. For institutional or national context, there should be a way for either having a plentiful of provenance information on the metadata itself in the metadata exchange, or by having a standard for organizations which to follow, when having their research output metadata be aggregated to European Publication Information Infrastructure. Both approached would need a strong cooperation with the data providers, be it a institutional or national system, to agree on specifics relating to metadata and information on its reporting and collecting to these source systems.

Infrastructure would also the increase in comparability of data by developing automated methods to restructure and reclassify data in a uniform way on the basis of the bibliographic metadata as well as information from external sources. 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 levell.

## Deliverable

1. An outline of implementing the research publication metadata transfer in European Publication Information Infrastructure
    a. Implementing in European Publication Information Infrastructure
    b. Interoperability Platform to Support Implementation

# Use of deliverables and future collaboration

The work and deliverables of this STSM will provide for many ongoing and future projects in the research domain relating to research metadata transfers and interoperability of databases at European level:

1. As a major input for the ENRESSH Working Group 3 objective to building up a database and design a roadmap for European database for SSH research outputs.
2. Presentation(s) in euroCRIS Membership Meeting in 2019 and/or euroCRIS Conference in 2020
3. Data model and implementation outline will support the application for Nordic Research Information System, a project proposal from 2018 by Nordic countries aimed to have a unified infrastructure for research outputs on Nordic level.
4. Connecting Europe Facility (CEF) Telecom funding call in later half of 2019 might be relevant for further funding on the European Publication Information Infrastructure (https://ec.europa.eu/inea/sites/inea/files/cef_telecom_work_programme_2019.pdf)
5. The findings provide valuable feedback for OpenAIRE and its current validation processes

## Supporting documents:

Puuska, Hanna-Mari; Guns, Raf; Pölönen, Janne; Sivertsen, Gunnar; Mañana-Rodríguez, Jorge; Engels, Tim (2018): Proof of Concept of a European database for social sciences and humanities publications: Description of the VIRTA-ENRESSH pilot. figshare. Journal contribution. https://doi.org/10.6084/m9.figshare.5993506.v1

https://dspacecris.eurocris.org/bitstream/11366/682/1/Puuska_et_al_CRIS2018_paper_Proof_of_concept_VIRTA-ENRESSH.pdf

Sle, L. et al. (2017). European Databases and Repositories for Social Sciences and Humanities Research Output. Antwerp: ECOOM & ENRESSH. https://doi.org/10.6084/m9.figshare.5172322

http://enressh.eu/wp-content/uploads/2017/09/2017_ENRESSH_European_Databases.pdf

Towards the integration of European research information

https://dspacecris.eurocris.org/handle/11366/593

https://openaire-guidelines-for-cris-managers.readthedocs.io/en/latest/index.html

CERIF-tietomallin määrittely OpenAIRE tiedonsiirrossa

CERIF - VIRTA mapping

https://docs.google.com/document/d/1Rm4OMOUf3JEti6aLmCrnSilX-sbutknFTR7njeItmBc/edit?usp=sharing