European Publication Information Infrastructure - metadata transfers in European context

Joonas Nikkanen - Project Manager - CSC IT Center for Science
orcid.org/0000-0002-5036-6444, linkedin.com/in/joonas-nikkanen
November 20th, 2019 - EuroCRIS Münster
1. BACKGROUND
   - Issues in data comparison of publications
   - ENRESSH point-of-view

2. PROPOSED INFRASTRUCTURE
   - STSM findings
   - Outline of infrastructure

3. WHAT’S NEXT?
Albeit European countries have invested heavily in the development of national research information infrastructures in the past years, there is not yet a way to do meaningful cross-country comparisons and international benchmarking for research publications across disciplines.
Well-structured and commensurate data

Publication/citation index databases
- e.g. Web of Science, Scopus, Academic

National CRISs or databases
- e.g. CRISTIN, VIRTA, VABB-SHW

Aggregating harvesters
- e.g. OpenAIRE, Google Scholar, Dimensions

Weak coverage
(in terms of disciplines and publication types)

Good coverage
(in terms of disciplines and publication types)

Less structured and miscellaneous data

Behind the three main types of data:

1. Background
   - National CRISs or databases
     - e.g. CRISTIN, VIRTA, VABB-SHW
   - Aggregating harvesters
     - e.g. OpenAIRE, Google Scholar, Dimensions

2. Proposed infrastructure
   - Good coverage
     - International
   - Weak coverage
     - National

3. What’s next?
The main objective of this Working Group is to reflect upon the standardization and the interoperability of current research information systems dedicated to the SSH research outcomes.

Task 3. Develop common rules and procedures for building the databases.

https://enressh.eu/
VIRTA-ENRESSH

Proof-of-Concept

https://wiki.eduuni.fi/x/X37qAg

https://doi.org/10.6084/m9.figshare.5993506.v1
Especially for SSH but not excluding other fields

Carried out between 3/2017-3/2018

Involved partners from Belgium, Finland, Norway, and Spain

Founded on the efforts made at national level in participating countries

The technical solution builds on the strengths of the Finnish VIRTA Publication Information Service
<table>
<thead>
<tr>
<th>Country</th>
<th>Category</th>
<th>Type</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finland/Madrid</td>
<td>Peer-reviewed articles</td>
<td>A1 Journal article, original research</td>
<td>VABB-1: journal article</td>
</tr>
<tr>
<td></td>
<td>Non-peer-reviewed articles</td>
<td>B1 Non-refereed journal articles</td>
<td>VABB-6: journal article</td>
</tr>
<tr>
<td></td>
<td>Monographs</td>
<td>C1 Book</td>
<td>VABB-2: monograph</td>
</tr>
<tr>
<td>Flanders</td>
<td>Peer-reviewed articles</td>
<td>A2 Review article</td>
<td>VABB-4: book chapter</td>
</tr>
<tr>
<td></td>
<td>Non-peer-reviewed articles</td>
<td>B2 Book section</td>
<td>VABB-4: book chapter</td>
</tr>
<tr>
<td></td>
<td>Monographs</td>
<td>C2 Edited book</td>
<td>VABB-3: edited book</td>
</tr>
<tr>
<td>Norway</td>
<td>Peer-reviewed articles</td>
<td>A3 Book section</td>
<td>VABB-5: proceedings paper</td>
</tr>
<tr>
<td></td>
<td>Non-peer-reviewed articles</td>
<td>B3 Non-refereed conference proceedings</td>
<td>VABB-5: proceedings paper</td>
</tr>
<tr>
<td></td>
<td>Monographs</td>
<td>C3 Textbook, professional manual or guide</td>
<td>VABB-3: edited book</td>
</tr>
<tr>
<td>Professional</td>
<td>D1 Article in a trade journal</td>
<td>D4 Development or research report</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D2 Article in a professional book</td>
<td>D5 Textbook, professional manual or guide</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D3 Professional conference proceedings</td>
<td>D6 Edited professional book</td>
<td></td>
</tr>
<tr>
<td>Popular</td>
<td>E1 Popularised article, newspaper article</td>
<td>E2 Popularised monograph</td>
<td>VABB-2: monograph</td>
</tr>
<tr>
<td></td>
<td>E3 Edited popular book</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Journal articles, all

0%  | 50%  | 100%  
---|---|---
Helsinki | | |
Jyväskylä | | |
Tampere Tech | | |
Oslo | | |
UC3M | | |
Antwerpen | | |

### Journal articles, identified as level 1-3

0%  | 20%  | 40%  | 60%  | 80%  | 100%  
---|---|---|---|---|---
Helsinki | | | | | |
Jyväskylä | | | | | |
Tampere Tech | | | | | |
Oslo | | | | | |
UC3M | | | | | |
Antwerpen | | | | | |
Issues in data comparability:

- Disciplines
- Inclusion criteria
- Semantics
- Publication types
Develop the existing VIRTA-ENRESSH data model by **taking into consideration the CERIF data model** and achieve wider interoperability by **forming a minimum set of CERIF elements needed in publication metadata transfers**

Investigate the potential of VIRTA-ENRESSH data model as a basis of research publication metadata transfers in European context and **outline how the research publication metadata could be transferred** by using the ENRESSH-VIRTA infrastructure as an example of national metadata aggregator

ENRESSH
Short Term Scientific Mission – June 2019
Data model

Set of CERIF elements with control over which attributes need to be included, divided into three categories:

**Mandatory**
- Publication
- Internal identifier
- Publication type
- Publication title
- Publication date
- Author
  - Organizational author and affiliation
- Discipline

**Conditional**
- ISSN*
- ISBN*
- Source title*
- Peer review*

**Optional**
- Audience
- DOI
- Volume
- Number
- Start page
- End page
Publication channel databases

To harmonize publication level metadata, a publication channel database (e.g. Publication Forum, Nordic list, ERIH+) and book publisher database (e.g. IRAP) could be used to resolve classification and publication type issues.

Also provides a qualitative aspect to individual publications and assists in identifying predatory journals.
Metadata transfers

Technical solution based on the ENRESSH-VIRTA-POC with addition of the use of (CERIF compatible) OAI-PMH endpoints

Accommodates to variety of source systems with varying technical capabilities
ENRESSH  
Short Term Scientific Mission – June 2019

Data model for European Publication Information Infrastructure

Following the ENRESSH-WIRTA-JIOC, 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 FOC of 6 organizations and 4 countries, a certain set of classes, attributes and associations were observed that could make for a so-called “forest 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 competitions and benchmarking.

As one deliverable of this STM 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 ontological-based.

Results:  
https://wiki.eduuni.fi/x/S4t_Bg

Outline of the architecture for European Publication Information Infrastructure

Results:  
https://wiki.eduuni.fi/x/TYt_Bg
Well-structured and commensurate data

Publication/citation index databases
e.g. Web of Science, Scopus, Academic

European Publication Information Infrastructure

Aggregating harvesters
e.g. OpenAIRE, Google Scholar, Dimensions

National CRISs or databases
e.g. CRISTIN, VIRTA, VABB-SHW

Less structured and miscellaneous data

Weak coverage (in terms of disciplines and publication types)

Good coverage (in terms of disciplines and publication types)

International

National
Emphasis on:

Controlled structure
Metadata quality
Inclusiveness
Comparability
ENRESSH coming to an end in April 2020

Work so far has been valuable and will continue in some shape or form

Collaboration in e.g. Nordic List development, NordRIS proposal for NordForsk

To continue work on the European infrastructure, an application for CEF-Telecom call was submitted last week
Joonas Nikkanen

Project Manager
Research Information Management and Interoperability
Tel. +358 50 381 80 92
linkedin.com/in/joonas-nikkanen

facebook.com/CSCfi
twitter.com/CSCfi
youtube.com/CSCfi
linkedin.com/company/csc--it-center-for-science
github.com/CSCfi