Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2016) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

13th International Conference on Current Research Information Systems, CRIS2016, 9-11 June 2016, Scotland, UK

Swepub Analysis - Offering High Quality Institutional Repository Publication Metadata Using Linked Data Technologies

Theodor Tolstoy*

National Library of Sweden, Kungl.Biblioteket Box 5039 102 41 Stockholm, Sweden

Abstract

SwePub is a bibliographic search service[†], harvesting and offering unified searching of aggregated scientific publication metadata from institutional repositories (IR:s) in Swedish universities and higher education institutions. SwePub has been developed by the National Library of Sweden.

Last year, in response to a government assignment, SwePub released a technical preview of an entirely new service – SwePub Analysis[‡] – aimed at researchers and analysts working in the areas of bibliometrics and scientometrics. SwePub Analysis is a bibliometric service enabling users to obtain enriched and validated scientific publication metadata to base their research and analyses on.

SwePub Analysis is built on linked data technologies and, together with data from other research information resources, allows users to query the database to obtain new knowledge concerning research information that would otherwise be difficult to obtain, e.g. richer Open Access information, deeper knowledge of scientific collaboration etcetera.

For the service to be able to provide high quality data, and for users to understand it's limitations, much effort has been spent on analysing and validating harvested metadata. This enables the service to present data providers with visualised, rich data on which elements are missing or do not meet format specifications and standards. Hopefully this approach will give IR:s incentives to improve data quality.

This paper outlines the present state of the service and planned development with emphasis on Swepub utilisation of linked data technologies and external data for validation and enrichment. It also contains insights on current developments in improving metadata markup of licenses and open access in order to improve Swedish Open Access statistics for the purposes of reporting.

* Corresponding author. Tel.: +46 70 716 10 67.
*E-mail address:*theodor.tolstoy@kb.se

[†] Swepub search web interface [http://swepub.kb.se/]

[‡] Swepub Analysis [http://bibliometri.swepub.kb.se]

Peer-review under responsibility of the Organizing Committee of CRIS2016.

Peer-review under responsibility of the Organizing Committee of CRIS2016.

Keywords: Type your keywords here, separated by semicolons ;

1. Introduction

1.1. SwePub's mission

is to enable and assure the quality of publication metadata for bibliometric analyses, a work undertaken in collaboration with the Swedish Research Council and universities and higher education institutions in Sweden. This mission was outlined in a government assignment issued in 2013.

1.2. The rationale behind Swepub Analysis

Swedish is a small language and, unlike English, not a common language shared by a large part of the scientific community. When looking at a graph showing the coverage and relation between Thomson Reuter's Web of Science (WoS) and SwePub this become obvious. Whereas WoS is the main choice for English STM publications, SwePub is more concerned with the humanities and social sciences, research areas which to a greater extent still publish in Swedish and in monographs. Using Swepub Analysis together with commercial services therefore provides a more complete picture of scientific research output in Sweden.

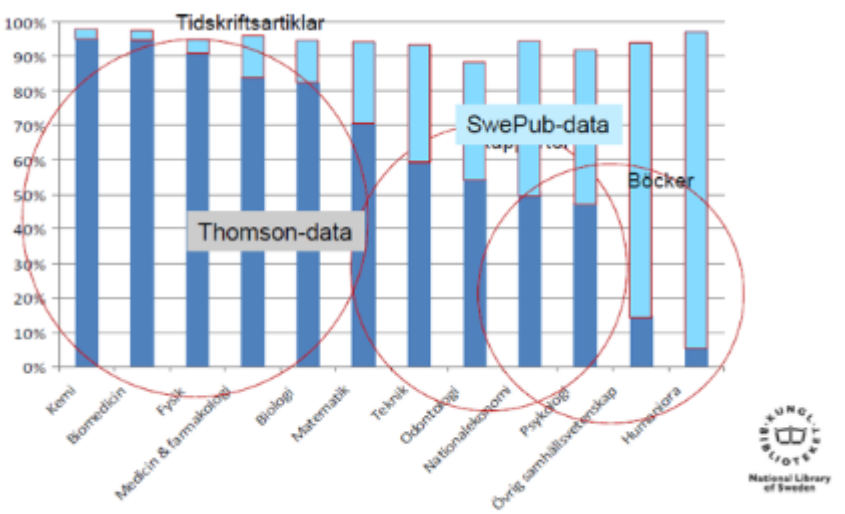

Fig.1. Data coverage in Swepub and ISI Web of Science

1.3. The role of Swepub in the Swedish research information ecosystem

The Swepub system as a whole plays an important role in the Swedish research information ecosystem. Swepub provides bibliographic data for many different systems, including the Swedish research grant application system

Prisma[§] , developed by The Swedish Research Council and jointly used by the funding agencies Formas, Forte and the Swedish Research Council.

Other important uses of SwePub Analysis are the quality assurance of publication metadata, research evaluation, co-publication studies, production volumes, publication trends, Open Access statistics etcetera.

2. The journey of data in Swepub Analysis

The SwePub system is a distributed system and publications are harvested, enriched, checked for inconsistencies, marked up with violation types, edited by local IRs, re-harvested in a cyclic process till the publication records are quality assured and meet the demands for bibliometric analyses.

Below is a brief discussion of this process.

2.1. Harvesting & triplification

IRs contributing metadata to Swepub do so in a Swedish profile of the MODS-format (Metadata Object Description Schema developed by the Library of Congress) through OAI-PMH-servers, and the records are then harvested by the system. Before they are stored in the analysis part of the SwePub system, publication records are converted into RDF triples.

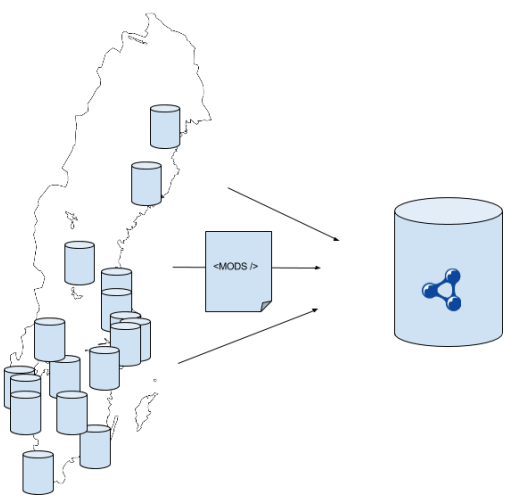

Fig. 2. Data harvesting into Swepub

Triplification can best be described as turning a human readable document into a machine readable graph. The graph allows one not only to store the data, it allows one to ask questions and reason about the data, and finally, it allows one to connect the data to other datasets on the internet.

Triple stores are built for importing and connecting data, and are very useful when it comes to validating data, or enriching with data from other sources.

2.2. Data validation

Data validation is a vital part of Swepub Analysis, and there are many good reasons for this. The primary reason is that we want to help IRs improve their data quality so that the overall quality of SwePub increases.

[§] Prisma [https://prisma.research.se/]

A second very important reason for our work in resolving data quality issues in Swepub Analysis, is that we want to give researchers or bibliometricians that use publication metadata a good idea of the validity of their research.

SwePub structure its validation process in three steps, through which a publication record has to pass before being considered as quality assured. The first step is concerned with making sure that harvested metadata is meeting the standards set out in the national metadata format specification. The second step is concerned with helping our deduplication algorithm to decide on which publications records are duplicates. The third step focuses on ambiguities in the metadata for those duplicate publication records that are the result of collaborations between researchers and/or universities.

The figure below is a screenshot from the web interface of SwePub Analysis showing a breakdown of various data quality issues for a specific university's publication records.

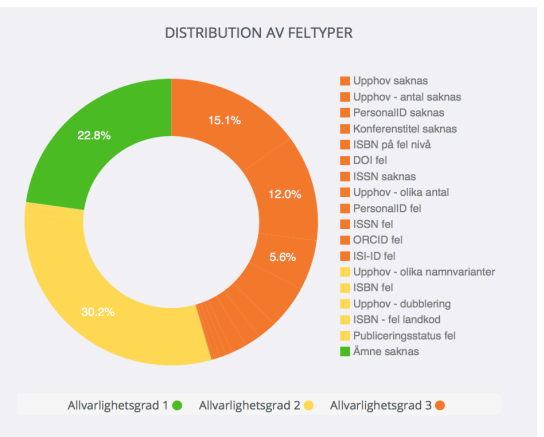

Fig. 3. Distribution of data quality issues for a specific university in Swepub

2.3. Enrichment

Enrichment is another important part of the system. SwePub analysis relies on external data not only for data validation, but also to be able to create new knowledge about data. The main focus regarding enrichment has so far been on publication channels. We have enriched our publication data with data from various existing nordic lists of scientific publication channels, like the Norwegian Register for Scientific Journals, Series and Publishers. We are also currently looking for new ways to validate Open Access publishing, something which is described later in the paper.

2.4. Data retrieval

There are numerous ways to retrieve data from Swepub. The underlying technique is using SPARQL-queries for this from the Swepub sparql endpoint[**]. Below is an example of a SPARQL query returning the number of theses in languages other than English and Swedish and getting results grouped by publication year and research subject.

[**] SPARQL endpoint for Swepub Analysis [http://virhp07.libris.kb.se/sparql]

```
PREFIX mods_m: <http://swepub.kb.se/mods/model#>
PREFIX swpa_m: <http://swepub.kb.se/SwePubAnalysis/model#>
SELECT DISTINCT
count(?IdentifiableRecord) as ?numberOfRecords
?_publicationYear
xsd:integer(?_hsv1) as ?HSV1
WHERE
{
?Subject a mods_m:Subject .
?IdentifiableRecord swpa_m:publicationTypeCode ?_publicationTypeCode .
?IdentifiableRecord mods_m:hasLanguage ?Language .
?Language mods_m:hasLanguageTerm ?LanguageTerm .
?LanguageTerm mods_m:languageTermValue ?_languageTermValue .
?IdentifiableRecord swpa_m:publicationYear ?_publicationYear .
?IdentifiableRecord mods_m:hasSubject ?Subject .
?Subject swpa_m:hsv1 ?_hsv1 .
FILTER(xsd:string(?_publicationTypeCode) ="dok") .
FILTER(xsd:string(?_languageTermValue) not in ("swe","eng")) . #For other languages
FILTER(xsd:integer(?_publicationYear) in (2011,2012,2013)) .
MINUS
{
?Language mods_m:objectPart ?_objectPart .
FILTER(?_objectPart IN ( "defence"^^xsd:string, "summary"^^xsd:string ))
}
}
GROUP BY ?_publicationYear ?_hsv1
ORDER BY ?_publicationYear ?_hsv1
```

Fig. 4. Example of a SPARQL query.

2.4.1. The web interface

For usability purposes though, SwePub has developed a web interface which handles generic data retrieval purposes. The interface is divided into two main workflows: one for bibliometricians and one for university staff working with data quality.

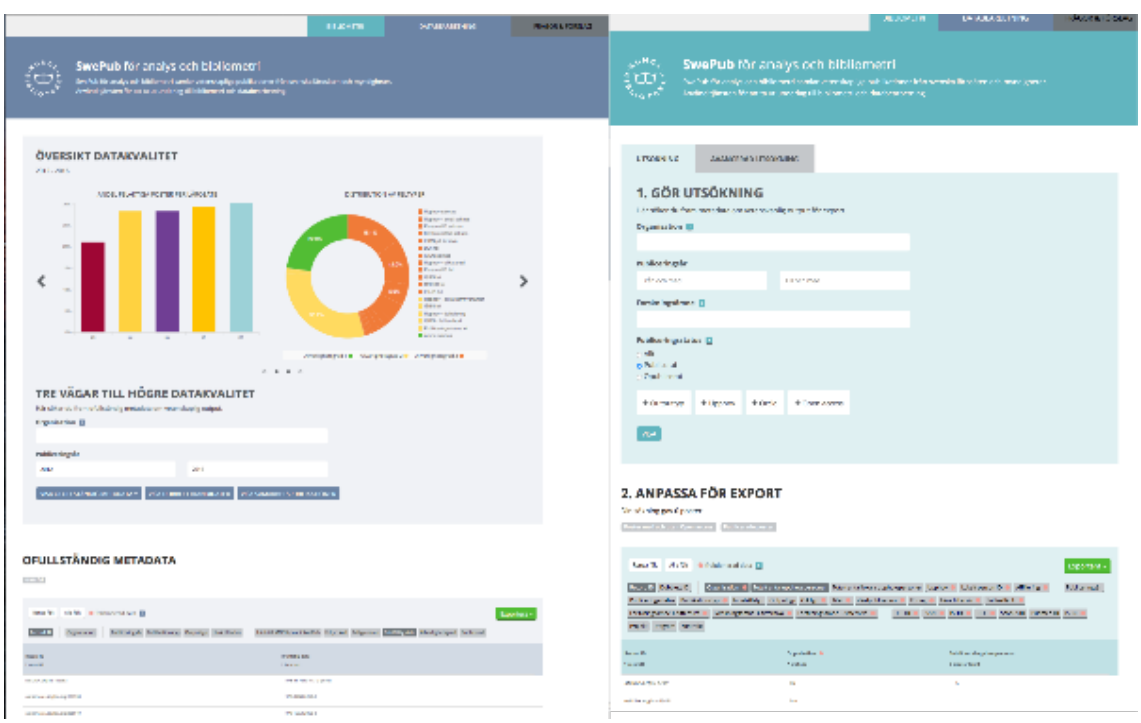

Fig. 5. Screenshots from the Swepub web interface.

2.4.2. The BI system

Shown below is a third option: a business intelligence system (Spotfire) where our users have the opportunity to make some basic analyses.

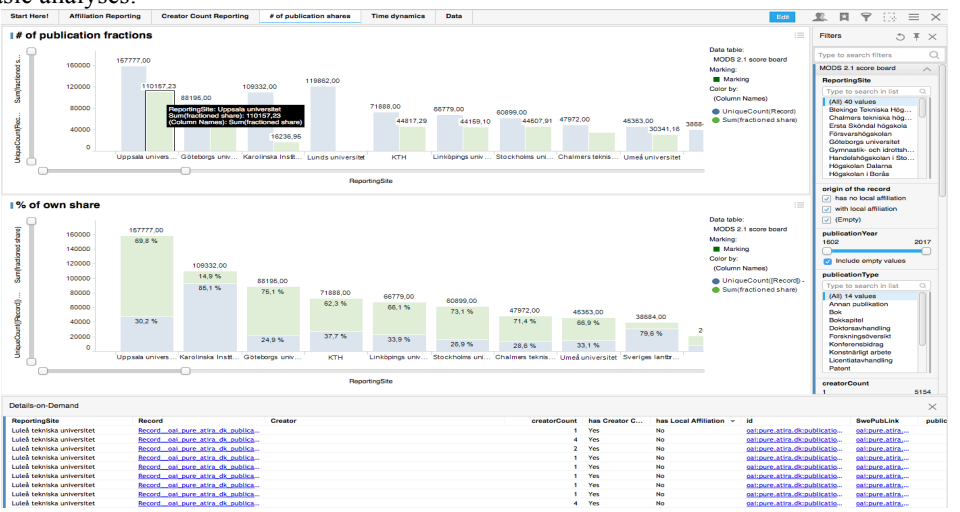

Fig. 6. Screenshot from Spotfire, the BI tool offered by Swepub

3. Open Access enrichment

Open Access enrichment is important for many reasons. One main reason is that many rely on SwePub to deliver data on the advancement of Open Access publishing in Sweden. Since universities do not always supply Open Access information themselves, e.g because of embargo times, we have to find other ways to provide this data.

Another reason is that SwePub can use the enrichment process to validate that publication records are published Open Access. Also, definitions on Open Access vary depending on whom you ask, so being able to filter out publications based on different Open Access criteria can be a valuable feature.

If we were to rely only upon data provided by local IRs for peer reviewed publications from the years 2010 through 2015, only 16% of the records would be registered as Open Access.

3.1. Directory of Open Access Journals (DOAJ)

We are currently relying on one external source for Open Access information: The Directory of Open Access Journals (DOAJ).

If we take into consideration publications that are published in journals covered by DOAJ, we can see (shown in the graphs below) that the number of Open Access published publications would rise to almost 25%.

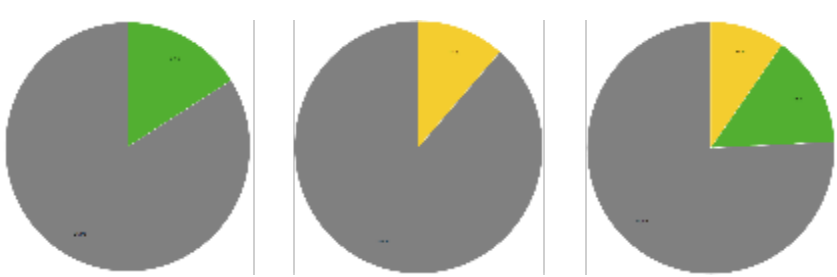

Fig. 7. OA coverage by type in Swepub

SwePub consider DOAJ a high quality source for Open Access information. They have high standards for accepting new journals and are also putting older journals through a reapplication process to be able to meet new and higher demands. There are currently 11 000+ journals in DOAJ.

DOAJ is also a rich source for other information. They provide, among other things, information on which licences publications use, and there is also information on article processing charges (APC).

There are a number of ways of retrieving data from DOAJ. You can either download an entire list of journals in CSV-format or use their API. They also provide data on article level.

3.2. ROAD - Directory of Open Access Scholarly Resources

ROAD is a service from the ISSN International Centre consisting of a subset of the ISSN Register where records have been marked as Open Access. ROADs criteria for selecting a publication channel differs from DOAJ both in types of resources and when it comes to determining the openness of a resource. ROAD contains not only journals, but also conference proceedings, academic repositories, book series and scholarly blogs. The majority of records refer to journals as is shown in the graph below.

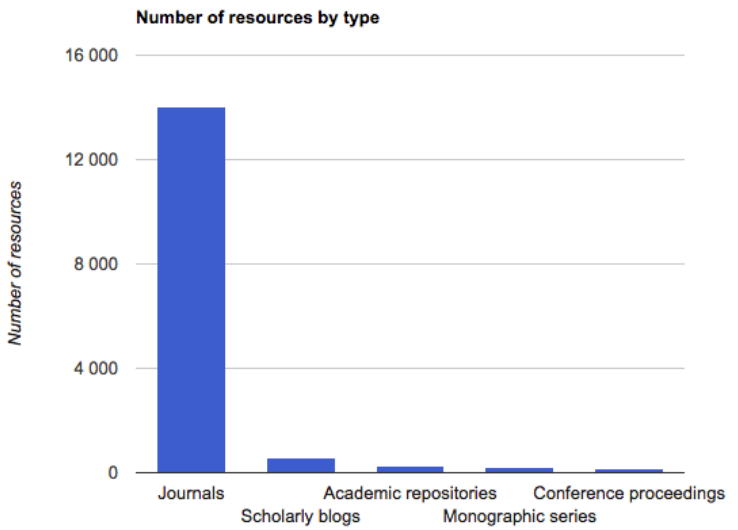

Fig. 8.Distribution of resources by type in ROAD. Graph generated from http://road.issn.org/en/statistics

The following Venn diagram shows how ROAD data overlaps with current data in SwePub. Most of the data is overlapping, but as the diagram shows, ROAD would certainly contribute to the picture of Open Access in Swepub.

|A|	21185
|B|	29708
|C|	40475
|A∩B|	18621
|A∩C|	8367
|B∩C|	11725
|A∩B∩C|	7521

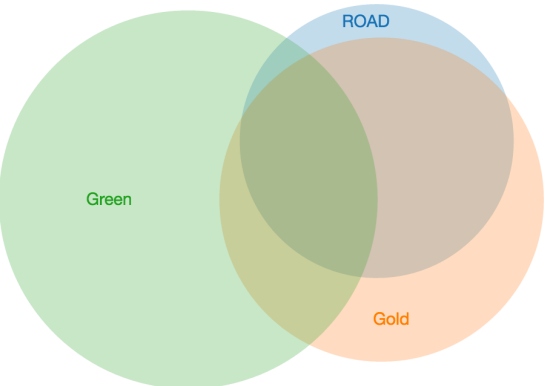

Fig. 9. Comparison of OA sources in Swepub and how they overlap.

The ROAD dataset also contains historic data on publication channels. Just like the ISSN registry itself, ROAD data conforms with ISSN cataloguing rules.

Integrating with ROAD data has been very straightforward since it is exposed as linked data in the format of RDF. MARC XML dumps are also available.

3.3. Directory of Open Source Books (DOAB)

Another source that SwePub is currently looking into is Directory of Open Source Books,DOAB.

DOAB currently contains more than 4000 academic peer reviewed books. A resource of this kind would give SwePub further insight to Open Access book, and book chapter, publishing. As of this writing, the amount of metadata on Open Access books is still small compared to the data coming from the local IRs contributing to SwePub.

There may be many reasons for this, one being the handling of ISBN:s in Swepub and in records supplied by local IRs. Since SwePub matches using ISBNs, we rely on submitted ISBNs being correctly registered. Another reason is the size and recent establishment of DOAB, and how Open Access book publishing differs from article publishing.

3.4. Collecting data on hybrid Open Access

When it comes to validating Open Access, data on hybrid articles is still difficult to collect. An article is denoted hybrid when an author has to pay an APC for having the article published Open Access in a traditional subscription journal.

There are currently a few initiatives in Europe collecting information on APC:s, and the Swepub project group is part of a pilot study, initiated by the Swedish Open Access Programme at the National Library of Sweden, that explores the possibility of collecting information on APC:s from different universities in Sweden. The pilot is inspired by work currently carried out in Germany called the Open APC-DE project[††]. Having APC data in Swepub Analysis would enable better Open Access reporting.

Another way of accessing information about hybrid Open Access publications is by asking the publishers themselves. Most publishers are able to provide lists on hybrid publications, but without enclosing the APC amount.

[††] The Open APC-DE Project [http://openapc.github.io/openapc-de/]

4. Future development

In Linked Data, the ontology is the data schema that makes it possible for data sets around the world to link to each other and making the data machine readable.

Swepub Analysis is a system that can consume linked open data, but Swepub Analysis mostly uses its own ontologies which makes it hard for other LOD systems to understand the data structure in Swepub. It is the project's ambition to change this, and we are investigating which ontologies to use and how to publish which data.

4.1. The Nordic list

Swepub - together with the Swedish Research council - represents Sweden in an initiative to develop a common Nordic list of authorized publication channels. Denmark, Norway and Finland are currently maintaining such lists on their own, and the initiative aims to provide higher quality and efficiency through collaborative work, shared definitions and a shared infrastructure.

4.2. Making Swepub Analysis fully operational

Swepub Analysis is in many ways already a fully featured system, but it is not yet fully operational and we are constantly working on improving and automating the different parts of the system.

5. Conclusions

Swepub Analysis is at an very interesting stage in its development. Most of its parts are fully functional and while helping the IRs on improving their data quality, the Swepub project can focus on improving interoperability with other systems and LOD datasets making it even more useful in the Swedish research information ecosystem.