Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2014) 000–000

Procedia
Computer Science

www.elsevier.com/locate/procedia

CRIS 2014

Showing it all – a new interface for finding all Norwegian research output

Nina Karlstrøm *, Lars Wenaas

CRIStin, University of Oslo, Norway

Abstract

CRIStin (Current Research Information System in Norway) is on the verge of implementing a SOLR based search engine, capable of indexing the entire national CRIS database as well as all the full text entries in NORA (Norwegian Open Research Archive). Perhaps the most useful feature of the search index, from a CRIS point of view, is as a tool for validating and improving data integrity. But in addition to this and the vastly increased speed, it will also provide the basis for a host of novel and exciting services like geographical mapping, timelines, RSS feeds and an externally available API.

Keywords: CRIS; SOLR; open access; indexing; institutional repositories; search engine

1. Introduction

The Norwegian national CRIS is currently being reimplemented on a new technical platform (CRIStin 2.0). One of the central services emerging is a new search engine based on the SOLR-engine combining the current two separate search indexes in the CRIS and NORA (Norwegian Open Research Archives). This is a strategic move to showcase Open Access publications to a wider audience through the CRIS and to thus strengthen both existing services and to lay a groundwork for a wide range of new services in the years to come. Although the search engine as such is ready for release, the CRIS itself is still under development; therefore the search engine will not be launched until the new CRIS is fully functional.

2. Background

The CRIStin organisation was established in January 2011 under the ownership of the Ministry of Education and

* Corresponding author. Tel.: +47-22852448;
E-mail address: nina.karlstrom@cristin.no

Research in cooperation with the Ministry of Health and Care services. CRIStin has three main functions: to manage and further develop the national CRIS system, to coordinate the implementation of Open Access in Norway and to negotiate licence agreements for e-resources on behalf of consortia of research institutions.

The CRIStin system is now in use in all Higher Education institutions, all hospitals and all research institutes receiving at least some degree of public funding. Most European countries have implemented institutional CRIS systems and collected information from these to a national level. CRIStin does it the other way around: CRIStin is the central data store from which data can be accessed by everyone. The institutions are also using the CRIS for uploading academic papers into their institutional repositories, making the connection between CRIStin and NORA even more obvious.

NORA (Norwegian Open Research Archives) is a search index containing content from all Norwegian institutional repositories. Approximately 80 000 documents are indexed in NORA (half of which are master theses). CRIStin contains about 1 000 000 records and ideally there should be full correspondence between all scientific entries in NORA and the CRIS, but in fact only a small number of the entries in NORA are identifiable with metadata from the CRIS.

The CRIS contains mainly bibliographical records of publications, not the full text itself. A successful marriage between NORA and the CRIS requires “integrational glue” of the two main sources of information on Open Access journals. DOAJ (Directory of Open Access Journals) is a database containing all Gold Open Access journals and Sherpa Romeo is a comprehensive database over scientific journals’ self-archiving rights granted by the publisher, so-called Green Open Access. Both these sources are integrated in the CRIS and are vital for colouring the CRIS with Open Access.

3. Goal

Since CRIStin is a complex system with over 1 million records, it is hardly surprising that we need a search engine. A web-based system of any magnitude or complexity needs a reliable way of finding and displaying data, and in this respect nothing new has been brought to the table. What we do focus on is the integration between CRIS data and full text-data, and the use of the search index throughout the system as a basis for other services.

The goals are as follows:

- A search engine that fulfills all requirements within the CRIS
- A tool for validating data posts and data integrity
- A flexible basis for new services in the years to come for both CRIStin and external partners

Choice of technology is important. SOLR is a professional and well documented search engine based on facets; SOLR is also a rock-solid open source project under the Apache foundation. SOLR gives a very fast response because it indexes all the words in the index whereas a standard RDBMS must run a search through all the documents to look up a specific word. Text-based databases like SOLR also compile statistics on the incidence of individual words in the whole text, thus yielding more comprehensive and relevant results.

The SOLR package gives us the opportunity to have a single common index as the foundation for multiple services. The search service in the main application (the CRIS) is of course the most important, but we are also launching two additional search services that will coexist with the main one: a service similar to the old NORA, and another that will only search master theses.

Flexibility is also important. At a later stage we can build new services, for example a search service exclusively for doctoral theses. The main object is therefore not only a search engine within the CRIS, though this is essential for most applications (and this is certainly challenging enough). The goal is also to build an index that serves as a foundation for other services within the CRIS. We would like to build a “superindex” that serves map services, timelines, tag clouds, web services and other forms of data integration etc. The index is, for example, well suited for a “build-your-own-RSS”-engine. Any keyword search and/or combination of facets can be expressed as a result delivered in the RSS-format, among others. Further into the future we are planning to incorporate a SPARQLendpoint and deliver data in RDF-format as well.)

The whole application can in principle be driven by the SOLRindex since the index actually contains all the data stored in the database. Any web page showing data from the CRIS can retrieve it directly from the index, rather than resorting to the slower process of looking it up in the RDBMS. (In fact the index serves as a metadata storage level on top of the database.)

Not only the CRIS-application itself can benefit from the index, but other applications on other sites can also access it. The possibilities for RSS streams have already been mentioned; additionally a university website can fetch data through an API, displaying for example an individual researcher's projects or publications.

The following illustration is a glimpse of how we see the development of services from the beginning with multiple CRISes, the unification of CRISes into one single instance, the merging with NORA and then into a future we can neither know nor should try to predict too closely.

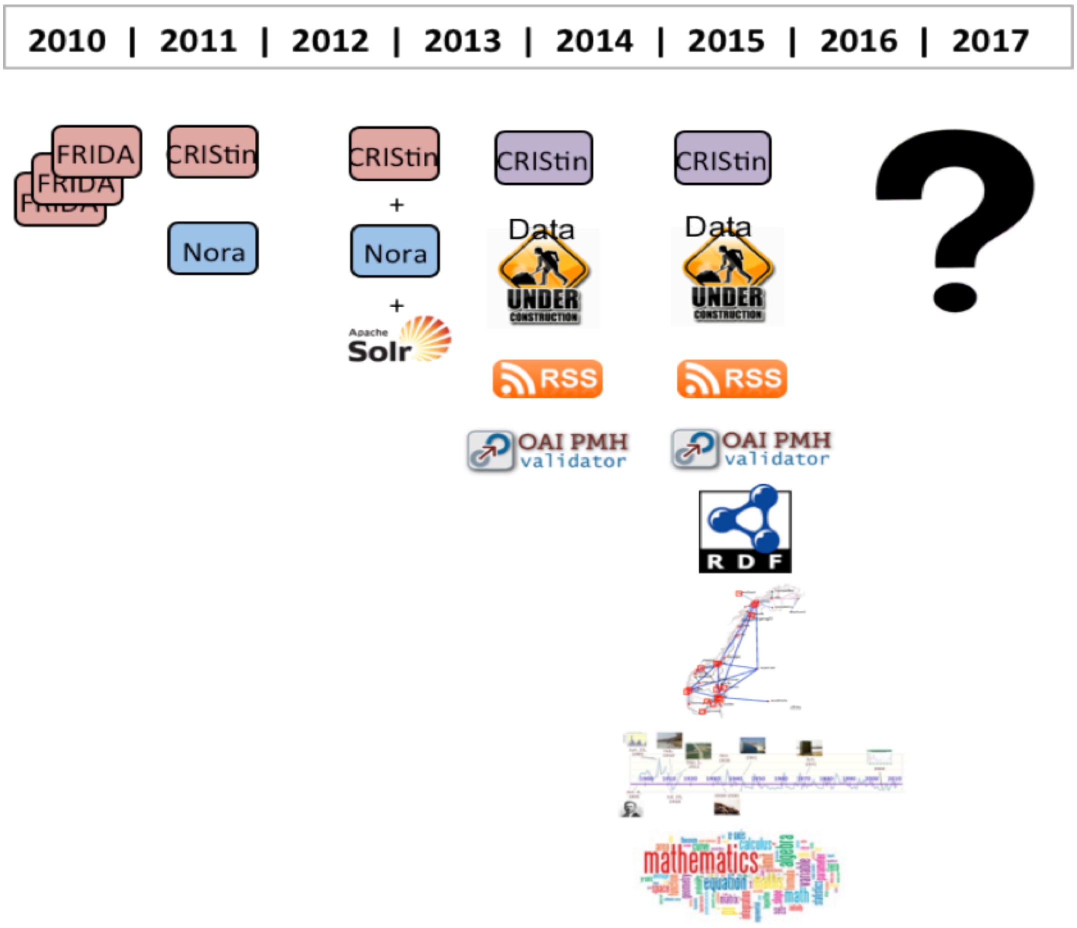

Figure 1: Development of search engine

Another vital goal for the search index stems from the fact that all data in the CRIS and all data in Norwegian repositories now passes through a single gate on its way into the index. This gives us the potential to validate every bit of data. Even though there are rules for data integrity in both the application layer and in the database layer, we can now check each post for missing or poor quality data as seen from the user perspective. The application itself may allow for missing keywords on a given record. From a searching-interface viewpoint this harms the searchability of data in general, and the post in particular.

4. Challenges

In an ideal world, all data wraps up just fine and corresponds the way we want it to. In the ideal world all the links and connections between the different entities in the data model above are populated with strict and coherent data. However, the ideal world hardly represents the real world. Our first discovery when we built the search engine is that data in the CRIS and in the repositories are lacking essential information to accommodate the search interface as we would like it to do. This makes the log server even more important. It will take a lot of time and a great effort to achieve acceptable data quality anywhere near the optimum. Typical missing data are thematic categorisation, keywords, corresponding authors, abstract, second language abstract and DOIs. These issues have to be addressed by all the CRIStin institutions.

5. Application buildup

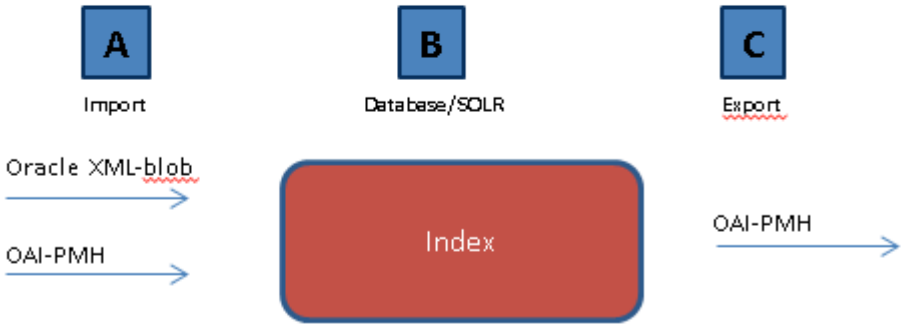

Figure 2: Buildup of application

A. Import of data

As shown in the illustration, there are two main sources for import. The first source is the data from the CRIS, mainly records of research results, but also data on projects, people and institutions. These are generated as XML-objects from Oracle. The XML is parsed through a validation layer to form the basis of the SOLR database.

The second source is the import of data from Norwegian institutional repositories. There are about 60 Norwegian repositories, all of them being treated one by one in parallel harvesting. The data contains metadata information on publications and links to the actual document. All Norwegian IRs are based on Dspace, a software compliant with the transmission protocol OAI-PMH.

The full text documents are mainly PDFs, but other formats may also occur. The harvested XML contains links to the PDFs which are downloaded and transformed to plain text (by the TIKA-software) and then indexed in SOLR.

A separate database serves as a temporary storage, making it possible to re-index all raw data from CRIStin and all content from the repositories.

Validation of records

Import of data will go through a separate validation process. This validation layer will check for errors which are logged to a separate SOLR-index where they can be further examined. This log server will perhaps prove just as important as the search service itself. Since the log server acts as the gatekeeper where all rules of data integrity is stored, it also signals to administrators that a particular post must/should/could be refined.

B. The SOLR database

The service uses SOLR to index metadata and text. The elements consist of:

- Research results
- Authors
- Projects
- Institutions
- Journals
- Research groups
- Publishers
- Institutional Repositories

All of these elements are searchable, and each entity will have facets on the other elements according to the simplified data model in the illustration. This means that one can easily narrow down a search - for example: find all publications the field of biology that are Open Access.

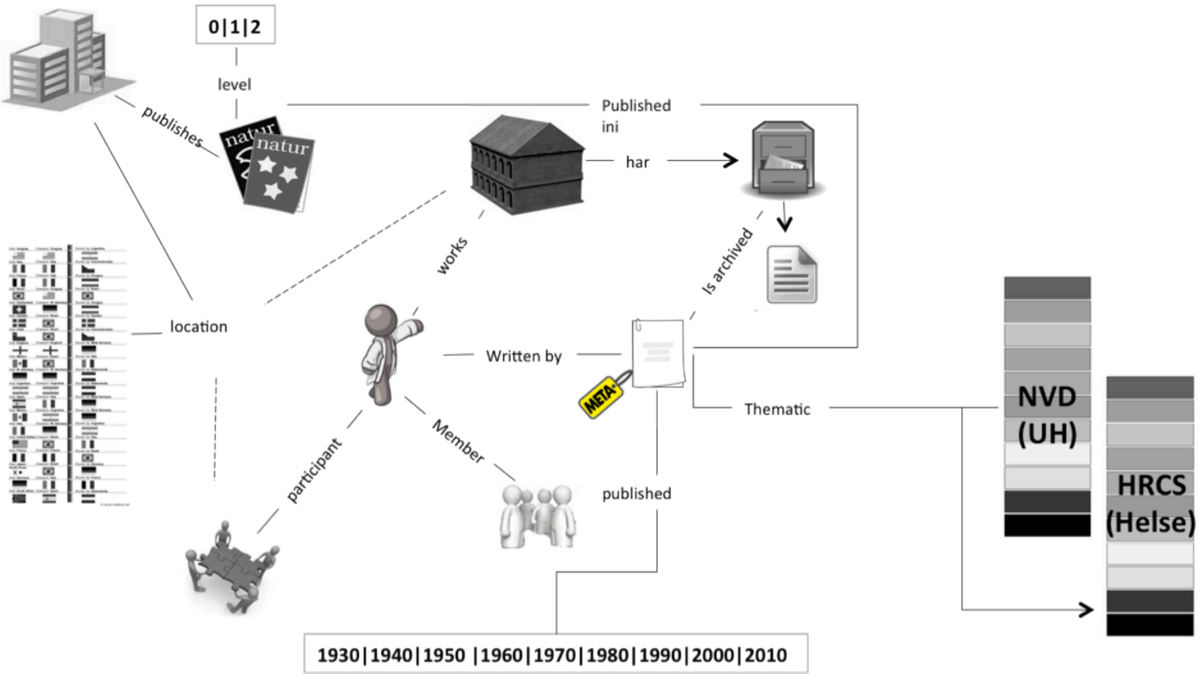

Figure 3: SOLR elements

Where publications are available in full text this will be indexed as well, completing the metadata record and enhancing the searchability of that particular record.

C. Export of data

One of the most important possibilities to arise from the SOLR-index is the provision of a service for the export of data in different formats. Any query can return any result on virtually any format, as long as it is supported by the data model.

From an Open Access view (i.e. those services concerning NORA) this possibility is the most important service of all. The main feature of NORA has always been to harvest content from the repositories and disseminate data to external services such as Google Scholar, services within libraries, commercial discovery services etc. The services that are or will be launched are based on formats like OAI/PMH, RSS and Json. Other export formats may be added at a later stage, the most interesting one being the CERIF-model, which makes it possible to communicate with other CRISes around the world. A full compatibility with CERIF is planned for the new CRIStin application.

6. Summary

We believe there is a great potential in bringing together the results in the CRIS with the full text documents in the IRs, making Norwegian research output more visible. A many-faceted search index will also provide a much-needed upgrade of the end-user side of the CRIS service. Furthermore, we need scalable tools to administer records in the CRIS. As an added bonus we see the potential of new services for the CRIS such as geographical mapping, timelines, RSS feeds and an externally available API. The achieved result is a better way of finding Norwegian research output, a means to showcase Norwegian the publications and a smooth dissemination to national and global services.