Akademisk forskning online - Academic Research in Sweden Online

A pilot study of an OAI compliant portal

Eva Müller
Uwe Klosa
Peter Hansson
Stefan Andersson
Uppsala University Library, Electronic Publishing Centre
eva.muller@ub.uu.se
uwe.klosa@ub.uu.se
peter.hansson@ub.uu.se
stefan.andersson@ub.uu.se
Box 510, 75 120 Uppsala, Sweden
publications.uu.se/
Keywords: Interoperability, OAI-PMH, DiVA, Metadata, resource discovery

Abstract

An increasing number of universities in Sweden make their university research results available in electronic form. Technical solutions for electronic publishing are different, but the aim is usually the same - to facilitate the efficient discovery and dissemination of content.

Five universities in Sweden have already agreed on a common XML schema for publishing electronic documents. The DiVA portal (http://publications.uu.se/portal/) is built on this common schema. The portal has two main functions:

1. Federated searching for theses, dissertations & other electronic publications at a number of Swedish universities

2. Metadata publishing service (the portal supports OAI-PMH-based metadata harvesting).

Within a pilot study funded by BIBSAM (the Royal Library´s Department for National Co-ordination and Development), we tried to find a solution that would allow other universities to take part in the common portal. That is, to make their research results, particularly theses and dissertations, available through a catalogue built by metadata harvesting using OAI-PMH. Some of the universities use very simple solutions or have no technical staff, making it difficult for them to implement an OAI data provider. In these cases, for which OAI harvesting is not practical, we tested alternative strategies to collect metadata.

To enable these archives to enhance access to their documents and increase the availability, we offered them to make their metadata available for OAI harvesting through the DiVA portal. In this way, even 'small' publishers could be part of, for example, OAI-based catalogue of ETD's. That we could reuse the technical solutions developed for the DiVA portal was one of our starting assumptions.

The result of this pilot study is a portal for federated searching of theses, dissertations, and other electronic publications from seven Swedish universities. Unfortunately, the quality of the harvested metadata does not allow reuse of DiVA portal components. Therefore, a simple search interface was developed. The inconsistent vocabularies and different interpretation of DC elements were the main problems.

This case study shows very clearly that common agreement on interpretation of metadata standards and vocabularies is necessary for meaningful resource discovery. Achieving a national SYSTEM that facilitates this access is a question of interoperability not only on the technical, but also on the content level (in terms of metadata and vocabularies used). This question of interoperability will be the focus of some new projects, submitted to BIBSAM for immediate decision on supporting grants.

Preface

Electronic publishing is a very good method for achieving efficient discovery and dissemination of content. An increasing number of universities in Sweden make their research results available in electronic form directly from locally hosted repositories. To date, Swedish universities have electronically published approximately 1300 doctoral theses, several thousand working papers[1] and about three thousand undergraduate dissertations. Some universities also publish their own electronic journals.[2]

Though the metadata for electronically published documents are usually a part of traditional library catalogues, the electronic publications themselves are often accessible directly through a various web services.

But what kind of added values could a Swedish portal for electronic published documents have? Do we really need a separate portal if all this information can be found in the national library catalogue and local catalogues? Is it realistic to build a shared portal for all Swedish universities that are active in electronic publishing of university material? What is the minimum interoperability level for meaningful resource discovery of academic publications published by Swedish universities?

These are some of the key questions that were explored within a pilot study, "Academic Research in Sweden Online," funded by the Royal Library´s Department for National Co-ordination and Development (BIBSAM).

1 A large number of these are a part of RePEc or other subject based services
2 For detailed information see the annex to the project report: *http://publications.uu.se/afo/afo_report.pdf*

Based on this pilot study, these questions will be answered from the perspective of finding a realistic, practical and functional solution that would allow Swedish universities to participate in a common portal that also exposes metadata for harvesting by other service builders.

Because electronic theses and dissertations (ETD's) are the most important part of our electronic production, we will pay particular attention to doctoral theses and dissertations in this paper. The first part of this paper describes some of the specific characteristics of doctoral theses in Sweden and how technical solutions can help to handle this in a rational workflow. After that we will discuss two portals solutions, one based on a federation strategy and the other one — developed for the pilot project — based on harvesting of metadata

Characteristics of (Electronic) Publishing of Doctoral Theses and Dissertations in Sweden

Doctoral theses in Sweden are traditionally published before the defense as books in printed form. Since the defense of all doctoral theses is public, rules require that theses be available for evaluation at least three weeks before the defense event. By extension, the electronic publication of theses in Sweden must also satisfy these conditions. In reality, this means that theses publishers face tremendous time pressure. At the same time, electronic publication creates the potential to develop new, innovating services. The DiVA [3] Publishing System was developed to fulfill these requirements. In addition, several new services were developed and integrated into the workflow.

DiVA Publishing System

Technical solutions for electronic publishing are various, but the aim is usually the same — to facilitate the efficient discovery and dissemination of content.

Some of the publishing SYSTEMs used by Swedish universities allow improvements in the publishing workflow. The DiVA System, developed by Electronic Publishing Centre at Uppsala University Library[4], is one such SYSTEM. This SYSTEM makes it possible to reuse and enhance the data originally entered by the author as the basis for creation of all metadata and the "digital master" for both the electronic and printed version of the document, to store these documents in the local depository, assign a persistent identifier, checksum the file and to send a copy of the document to the Royal Library[5] (Swedish National Library) to support long-term preservation in the National Library Archive.

In addition to theses and dissertations, the DiVA document model supports a number of other document types. Currently there are five universities in Sweden using this SYSTEM[6], but the number is increasing. All DiVA cooperation partners support common technical, content and organizational agreements, so the level of interoperability is very high. This strategy makes it possible to develop high quality collections and services at a reasonable cost.

It also makes it easier to create new innovative — or even experimental —services.

DiVA-portal Example of a portal building on the federation strategy

One of new services we would like to mention here is DiVA Portal.[7] Because all of the participants support a number of agreements, as for example the common document format - DiVA Document Format,[8] is the interoperability very high.

Another of the agreements within the participants is for example to disseminate metadata in many different formats through the DiVA Portal, instead of directly from the SYSTEMs of participating universities.

The strategy used here for dissemination of metadata from a single point makes the portal more attractive for service providers and at the same time guarantees that all participants' metadata are disseminated in a unified format.

It also makes it easier to keep track of format standards and technology changes and to implement these updates in a timely manner.

The technical solution is scalable, thus new participants can be easily added to the service.

The content of the DiVA Portal is built on the DiVA Document Format. The portal has two main functions:

3 DiVA - Digitala Vetenskapliga Arkivet
4 *http://publications.uu.se/epcentre/index.xsql?lang=en*
5 *http://www.kb.se/*
6 Within the DiVA project, originated at the Uppsala University, five Swedish universities cooperate. The participants are the universities of Stockholm, Södertörn, Umeå, Uppsala and Örebro.
7 *http://publications.uu.se/portal/*
8 *http://publications.uu.se/schema/1.0/diva.xsd*

Federated searching for theses, dissertations and the other electronic publications of a number of Swedish universities

Metadata publishing service (the portal supports OAI-PMH[9]-based metadata harvesting. Using dynamic mappings directly from the DiVA Document Format a number of metadata formats are produced).

From the point of architecture both DiVA Publishing System and DiVA Portal are built by using methodology of component-based design. That allows modularity and reusability of the components. In addition, modules can be seamlessly replaced with improved implementations of the component.

The Pilot Study - Academic Research in Sweden Online Example of the portal build by harvesting

Within a pilot study funded by BIBSAM, we tried to find a solution that would allow other universities— not just those participating in the DiVA Portal — to take part in the common portal. That is, to make the participants' research results, particularly theses and dissertations, available through a new web-based service.

The suggested solutions would meet some conditions. Some of them were an easy integration of participating collections and low barrier technical solution.

For metadata transmission we thought it was realistic, practical and functional to use the Open Archive Initiative protocol for harvesting of metadata (OAI-PMH). Our assumption was that even small collection builders could adopt the OAI-PMH protocol quickly and relatively easily. The protocol addresses both interoperability and extensibility. Extensibility makes it possible to transmit even community specific metadata. In the context of Swedish theses and dissertations, for example date and place of public defense of theses is such a metadata. This metadata could be used for building new services such as, for example, a service to advertise upcoming public defenses or forthcoming theses on the national level. Such a service is coveted, not only by researcher themselves, but also by the Swedish public.

As a technical framework for the service, we wanted to reuse components of the DiVA Portal.

During our study, however, we found that using OAI-PMH for metadata transmission was not, in fact, immediately applicable for all repositories.

Six of Swedish repositories could be harvested directly (five DiVA Portal participants and another repository using E-Prints software).

But some of the universities use very simple, sometimes not even database based, solutions or have no technical staff, making it difficult or impractical for them to implement an OAI data provider. In these cases, we successfully tested alternative strategies to collect metadata (mapping to a XML schema and harvesting).

We discovered that these alternative methods are powerful and easy to implement, as long as structured data exists. It does, though, still require agreements on the organizational and content level.

At present, a number of Swedish universities are interested in using a publishing SYSTEM to improve their own publishing workflow. Because most available SYSTEMs allow control of the basic metadata granularity and are OAI compliant already, chances are good that the question of the basic technical interoperability for metadata transmission will be solved in the near future.

To make all Swedish repositories OAI compliant faster, we suggest an interim solution where alternative methods for collecting of metadata could be used. To enable these repositories to enhance access to their documents and increase the availability, we have offered to make their metadata available for OAI harvesting through the DiVA Portal. In this way, even those universities that decide not to implement an OAI data provider locally could be part of, for example, an OAI-based catalogue of ETD's.

Reuse of the technical solutions developed for the DiVA Portal was one of our starting assumptions. The DiVA Portal architecture, which separates metadata from presentation using XML-XSLT technology, makes it very easy to reuse the technology with a completely different design. Additionally there are modules supporting interoperability on the technical level - a separate harvesting module and a module supporting OAI - data providing - and modules for content interoperability providing metadata transformation to many formats.

Unfortunately, because of the low quality of the harvested metadata, it makes little sense to reuse DiVA Portal components.

Therefore, a simple search interface was developed. Only a few data elements are searchable in a structured form. Lack of content interoperability -inconsistent vocabularies and different interpretations of DC elements — was the main problem. The granularity of the descriptive metadata in the original collections was another problem.

Taking existing implementations of metadata schemes, we could recognize that, in many cases, only a subset of some standard was used, usually some basic elements of unqualified DC. The implementation of metadata in these services was often very pragmatic and, in many cases, directly related to the need to produce a simple search interface or records listing. Additionally, some-

9 Open Archive Initiative protocol for harvesting of metadata (OAI-PMH) *http://www.openarchives.org/*

times site-specific metadata was added to suit some local purpose,.

This restricts an aggregator service to only a very basic level of collections integration.

In the interest of solving several of these problems, we suggest an interim solution based on the ability to collect metadata using the alternative methods described previously. To solve some content interoperability problems, especially on the vocabulary and granularity level, it would be possible to map the specific vocabularies of each of the collections to the unifying vocabulary of the collector. The Digital Library Group[10] at Uppsala University will investigate this method and develop some tools within another project - called Personalize Access to Distributed Learning Repositories (PADLR).[11]

The result of this pilot study is an experimental portal for searching of theses, dissertations, and other electronic publications from seven Swedish universities built by harvesting. Because of the quality of harvested metadata the functionality is very basic.

This case study shows very clearly that common agreement on interpretation of metadata standards and vocabularies is necessary for meaningful resource discovery.

Achieving a national SYSTEM that facilitates this access is a question of interoperability not only on the technical, but also on the content and organizational level.

What is the Necessary Interoperability Level for Meaningful Resource Discovery of Academic Publications Published by Swedish Universities?

Within the pilot project we did a comparison between two portals - one built on the federation strategy with agreements on the technology, content (in terms of metadata and vocabularies used) and organizational levels (sharing costs, agreements about access to data, joint new development planning), and the other one based on harvesting of available metadata.

It was clear that interoperability strategies have a direct impact on the quality of service and results delivered. The high interoperability level among participating institutions in DiVA Portal makes it possible reach a high level of granularity resource discovery and building of new innovating services.

On the other hand, it makes it difficult to add services that don't support the same level of content interoperability (agreements on metadata format and semantic interpretation) as existing participants do.

However, there is a possibility to integrate other resources, for example harvested metadata by OAI-PMH, in a simple search interface and allow resource discovery at a more basic level. But how useful is this?

Conclusions: Levels of Interoperability = Granularity of Resource Discovery and Services Enable. Does It Make Sense?

What can all this tell us about the role of new services in comparison with traditional library catalogues? What kind of added values this new innovated services has?

What is necessary interoperability level for meaningful resource discovery of academic publications published by Swedish universities?

As we could see in the example of the DiVA portal the level of interoperability has a crucial role for the possibilities to offer new services and tailored community specific services. This is a great advantage in comparison with traditional library catalogues. The success of the DiVA Publishing System and the DiVA Portal is not measured only by the improved workflow, in which, for example, the web service "feeds" the library catalogue with basic metadata, but also by other aspects like usability of the service, number of metadata formats disseminated and ability to create new, personalized services. In contrast to the traditional library catalogue, new services are more easily developed (the posting of forthcoming theses being one of the more important of these).[12]

Support for various interoperability frameworks, like OAI technical framework, is also one of important added values.

The case study shows how important the questions of interoperability are for resource discovery.

Hopefully these questions will be the focus of some new projects supported by BIBSAM in the near future.

The specific goals of these new projects could be

- To define and agree on different levels of content interoperability
- To build up a technical infrastructure supporting interoperability
- (illustration: OAI-PMH as a part of local publishing SYSTEMs, a technical framework build on RDF or on using a common XML schema)

10 A research group based at Electronic Publishing Centre and Department of Information Technology at Uppsala University

11 *http://projekte.learninglab.uni-hannover.de/pub/bscw.cgi/0/44911*

12 For example see: *http://publications.uu.se/theses/*

- To achieve some basic organizational interoperability agreements on exchanging of data, rules for access and reuse of data.

1. Low barrier interoperability level (illustration: simple DC for using within OAI technical framework)
2. A richer interoperability level (illustration: URN: NBN; MARC XML; DiVA XML supporting long term preservation)
3. Interoperability on vocabulary level (controlled vocabularies)