

Title: Archiving Workflow between a local Repository and the National Archive

Experiences from the DiVA project

Authors: Eva Müller, Uwe Klosa, Peter Hansson, Stefan Andersson
Affiliation: Electronic Publishing Centre, Uppsala University Library, Sweden
Contact: Eva Müller, eva.muller@ub.uu.se, tel. +46(0)18 471 71 05

Content

Abstract.....	2
Introduction.....	4
DiVA Archiving Workflow	5
DiVA Workflow and DiVA Document Format.....	5
URN:NBN and Its Role in the DiVA Archiving Workflow	7
An Automatically Created Catalogue Record Including URN:NBN.....	8
An Archiving Workflow between the Diva Archive and the National Library Archive	9
Conclusions and future work	10

Abstract

In September 2000, an Electronic Publishing Centre was established at the Uppsala University Library. Its primary assignment was a project – DiVA project (Digitala Vetenskapliga Arkivet or Digital Scientific Archive) – in which technical solutions and a well-functioning workflow would support electronic posting and full-text publication of doctoral theses, essays, working papers and other types of scientific publications.

The long-term preservation of digital objects includes a variety of challenges. One of the objectives of the DiVA project has been to explore ways to ensure the future use and understanding of the digital objects in the archive.

Long-term preservation of digital objects is more useful if several copies of each document exist, preferably placed in multiple archives at multiple locations. It is also necessary to assign a persistent and unique identifier to each document. Therefore several projects were initiated in cooperation with the Royal Library in Stockholm.

A workflow and technical solutions supporting long-term preservation were created and implemented on both the local and national archive level.

The possibility of using XML as a format for long-term preservation was evaluated within the project. The DiVA Document Format – defined by an XML schema – has been developed. Using the DiVA Document Format for content management and inter-process communication, several applications were developed. Some of these provide services which are essential for long-term preservation:

- Make persistent National Bibliographic Numbers (NBN) available for the URN resolution service at the Royal Library in Stockholm
- Send MARC 21 records in MARC-XML to the National Library (including fields for electronic location and access with corresponding URN:NBN)
- Create archival file packages for long-term preservation, checksum them, store them in the DiVA Archive and send a copy of them to the Swedish Royal Library

This paper discusses the workflow, which was established between the DiVA-Archive and the Swedish National Library Archive. The experiences with this workflow are now the basis for a new proposal submitted to the Royal Library's Department for National Co-ordination and Development (BIBSAM) focusing on generalization of solutions developed within the DiVA-project. The objective of this new project is the development of a general archiving workflow between a local repository and a national archive.

Introduction

DiVA¹ – Digitala vetenskapliga arkivet (DiVA Archive) – is a comprehensive description of a searchable archive containing the documents, which are published in an electronic format at Uppsala University in Sweden.

The DiVA System, developed by the Electronic Publishing Centre at Uppsala University Library², makes it possible to reuse and enhance 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 (DiVA Archive), assign a persistent identifier (URN:NBN), checksum the file and to send a copy of the documents in an archival package to the Royal Library (Swedish National Library) to support long-term preservation in the National Library Archive.

This paper will discuss the workflow which was established between the DiVA archive and the Swedish National Library Archive to support long term preservation and future access to the digitally published documents and, more specifically, the significant role of URN:NBN (Uniform Resource Name:National Bibliographic Number) as an identifier in different parts of this process.

The results presented here have been achieved in cooperation with the Royal Library in Stockholm – the Swedish National Library³.

¹ <http://publications.uu.se/theses/index.xsql?lang=sv>

² <http://publications.uu.se/epcentre/>

³ <http://www.kb.se/>

DiVA Archiving Workflow

There is no guarantee that it will be possible to use and to understand digital objects in the distant future, but there are ways to increase the likelihood of success.

This assumption was the starting point for the discussions about the design of the DiVA Archive and the DiVA Archiving Workflow. We tried to find a practical and convenient way to minimize risks for data loss, both in the context of migration of the entire document and the connected metadata to other formats and media, but also in the context of accessibility.

The first step in this process was the decision to use XML as a primary storage format. The DiVA Document Format, with support for long-term preservation, was developed and the first fundamental steps of the construction of an archive for long-term preservation have been taken.

The question: “How can we ensure accessibility in the future?” initiated a couple of projects in cooperation with the Royal Library in Stockholm.

The usage of URN:NBN as a unique identifier and the exchange of metadata and archive-packages with the National Library Archive, were the next important steps in the development of this workflow. The DiVA archive and the archiving workflow refer to Open Archives Information System Reference Model.

DiVA Workflow and DiVA Document Format

The DiVA System was developed with a focus on how to achieve a rational and convenient workflow -- both authors and administrative staff -- for publishing electronic documents. The resulting workflow is based on the reuse of data directly from the documents originally created by the authors.

Templates, partly form-based, for MS Word, Open Office, Star Office, and LaTeX have been developed for use by authors when creating documents. To assure high quality metadata several controls are added to these templates. The templates produce XML files that contain all metadata and, in some cases, even the content of the document.

With XSLT and XSL:FO these files are used in various contexts to extract and to present the metadata and content.

After evaluating a variety of other formats, it was determined that none of them met the needs of the project completely. Therefore, it was necessary to develop our own format, which we call the DiVA Document Format. This internal format includes all the features that are needed to support the DiVA Workflow, the DiVA Publishing System and the DiVA Archive. Version 1.0 of the format, described in XML Schema, is component based and extensible. It consists of 99 elements.⁴ Administrative elements are combined with descriptive elements to make it possible to describe the publication within the same XML document file that contains the content. The DocBook DTD is used for the content part of the document.

Each item in the DiVA Archive is assigned a unique, persistent Uniform Resource Name and National Bibliographic Number (URN:NBN). As one of the identifiers in the DiVA Document Format the URN:NBN is used. It is used to map electronic resources to URLs and as a primary key of the publications stored in the DiVA Archive.

DiVA metadata can be mapped to a variety of metadata formats to support data reuse (oai_dc, MARC 21, etc).

⁴ See: <http://publications.uu.se/schema/1.0/diva.xsd>

URN:NBN and Its Role in the DiVA Archiving Workflow

A uniform resource name (URN) is a unique and persistent identifier for electronic resources on the Internet. The Royal Library assigns URNs in the Swedish national bibliographic number domain (urn:nbn:se) to organisations and the public in Sweden. The URN:NBN has several roles in the DiVA Archiving Workflow. The URN:NBN is also currently used to build the structure of the DiVA Archive⁵. In addition, the URN:NBN is used as the naming system of packages which are transmitted to the Royal Library.

DiVA uses URN:NBN primary as the identifier for each item – a single publication without consideration of its format (i.e., both the electronic and printed versions of a single thesis can be referenced by the same URN:NBN). Currently this is important because the possibility of multiple manifestations of the same item. The need for migration to new formats and new technology in the future make this characteristic even more important. By using URN:NBN it is also possible to point to more than one copy of the same item, which is essential for mirrored services, where one service backs up another. The distribution of the copies of the items to several distant places increases the accessibility. From the standpoint of long-term preservation, this identifier makes it possible for the archive to provide access to a document even after a catastrophic failure of the repository that originally provided that document. In our case, the Royal Library will make the archival copy accessible only in the event of a permanent shutdown of the DiVA Archive.

To resolve a URN:NBN, a resolution service exists at the Royal Library in Stockholm. Currently, the service has only a basic functionality, but there are plans for further development. A network of URN-servers, like the

⁵ See: Eva Müller, Uwe Klosa, Peter Hansson, Stefan Andersson, Erik Siira. Using XML for long-term Preservation : Experiences from the DiVA Project. ETD 2003 in Berlin. <http://publications.uu.se/etd2003/papers/LongTermPreservation.pdf>

network of DNS-servers essential for today's Internet, will be necessary to support international cooperation.

An Automatically Created Catalogue Record Including URN:NBN

The domain urn:nbn:se is managed by the Royal Library in Stockholm⁶. The Royal Library assigns sub domains to individual universities by appending their acronyms (e.g. urn:nbn:se:uu for Uppsala University). The DiVA system creates these URN:NBNs automatically, according the structure urn:nbn:se:<university>:<local-identifier>, as illustrated in the following example: URN:NBN:se:uu:diva-2154, where "uu" is the domain for Uppsala University and "diva-2154" is an identifier understood by Uppsala University. In the DiVA Archive, "diva" is the name of the archive / repository and 2154 is a serial number of an item within that archive. To refresh the information in the resolution service at the Royal Library, URN:NBNs from the DiVA Archive are harvested on a regular basis. Additionally, at the date of publication the DiVA-system creates a catalogue record in MARC-21 format, including the field URN:NBN, for inclusion in the National Library Catalogue, LIBRIS⁷.

⁶ See: <http://www.kb.se/urn/>

⁷ See: Eva Müller, Stefan Andersson, Uwe Klosa, Peter Hansson. Metadata Workflow Based on Reuse of Original Data. ETD 2003 in Berlin.
<http://publications.uu.se/etd2003/papers/MetadataWorkflow.pdf>

An Archiving Workflow between the DiVA Archive and the National Library Archive

In the DiVA Archive, all administrative and descriptive metadata are stored in XML that conforms to the DiVA Document Format. Today, the content of each document is stored in PDF and – where possible – in XML. Each document can be stored in different manifestations representing the same document in different formats (for example XML, PDF). On the date of publication, checksums for the files in the related manifestation are built and the metadata are locked and stored in the folder of the corresponding manifestation. After a certain time (currently 2 months) from the publication date, each manifestation is bundled into an archival package. Besides the manifestation file(s), typically a package contains general data and format specific data. General data is the XML document in DiVA Document Format (with all metadata and as much of content as we can currently get in XML) and the XML schema, which expresses the DiVA Document format at time of packaging. Example of format specific data is the style sheets, which can be used to extract the content of the XML files. This package is check summed and stored in the local archive for long-term preservation. This package and its checksum are then sent to the National Library Archive. Within a distributed system for long-term preservation, it would be possible to send these packages to multiple archives. This would significantly minimize the risk of data loss.

Conclusions and future work

The archiving workflow and the DiVA Archive presented in this paper are examples of a practical solution how the access to the content of the institutional research community outcome could be assured in cooperation with national libraries.

Our experience with this workflow is now the basis for a new proposal submitted to the Royal Library's department for National Coordination and Development (BIBSAM). Within this new project we will examine and evaluate current solutions. The project will focus on development and practical implementation of a generalized archiving workflow between a local repository and a national archive, focusing on the variety of publishing platforms and systems currently used by Swedish universities. Some other questions like for example - What is a minimal level of preservation metadata - will be explored. The project will also look at some other standards (f. ex. METS) as a possibility for packaging. Because of a lack of practical examples of implementations within the library community, we believe this project will be broadly useful.