

Archiving Workflow between a Local Repository and the National Archive

Experiences from the DiVA Project

Eva Müller, Uwe Klosa, Peter Hansson, Stefan Andersson

Electronic Publishing Centre, Uppsala University Library, Sweden

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

{eva.muller, uwe.klosa, peter.hansson, stefan.andersson}@ub.uu.se

Abstract. DiVA – 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 Library, 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 to the Royal Library, the National Library of Sweden, to deliver a deposit copy to the National Library Archive.

This paper discuss the workflow which has been 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.

Introduction

In September 2000, an Electronic Publishing Centre was established at the Uppsala University Library [4]. Its primary assignment was a project – DiVA project (Digitala Vetenskapliga Arkivet or Digital Scientific Archive) [11] – 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.

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.

The long-term preservation of digital objects includes a variety of challenges.

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, the National Library of Sweden [5].

A workflow and technical solutions supporting delivery of a deposit copy of electronic documents and long-term preservation were created and partly 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 [3]. Using the DiVA Document Format for content management and inter-process communication [8, 9], 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
- Send MARC 21 records in MARC-XML to the Royal 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 Long-term Archive and send a copy of them to the Royal Library.¹

This paper discusses the workflow, which has been established between the DiVA-Archive and the Swedish National Library Archive. The experiences with this workflow are now the basis for a new project funded by the Royal Library's Department for National Co-ordination and Development (BIBSAM) focusing on evaluation and 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.

The results presented here have been achieved in cooperation with the Royal Library [5].

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.

¹ This part of the workflow is not completely implemented yet.

The usage of URN:NBN as an unique identifier and the exchange of metadata and archive-packages with the Swedish National Library Archive, are the next important steps in the development of this workflow.

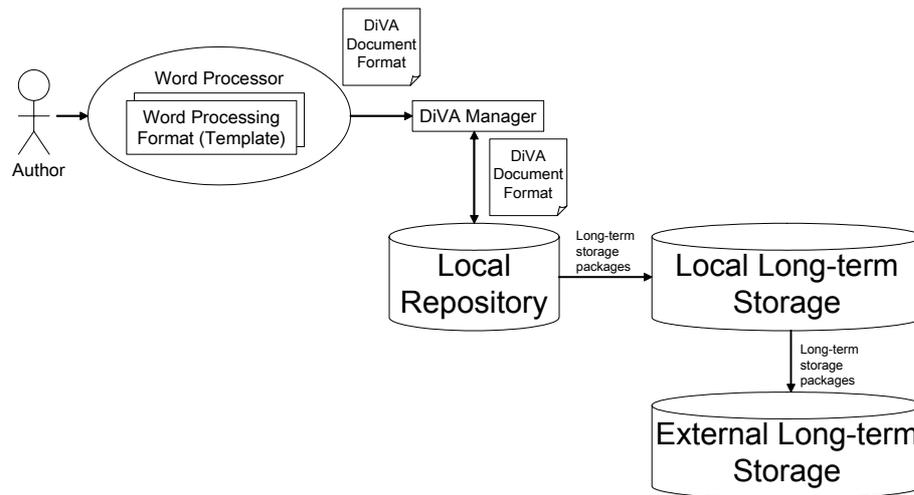


Fig. 1. DiVA archiving workflow

DiVA Workflow and DiVA Document Format

The DiVA System was developed with a focus on how to achieve a rational and convenient workflow - for 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 [3]. 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. If the entire document is stored in XML, the DocBook DTD is used for the content part of the document.

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

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

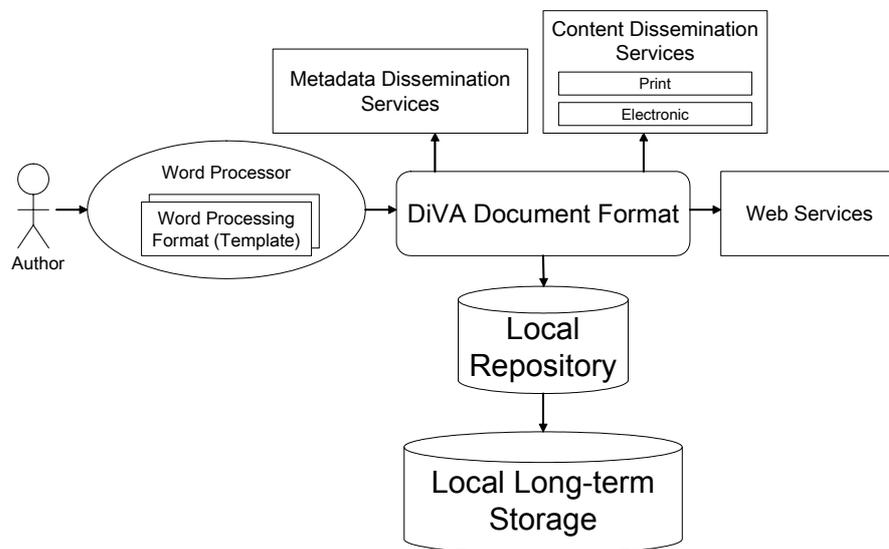


Fig. 2. DiVA workflow and DiVA Document Format

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.² Unlike a uniform resource locator (URL) a URN is a permanent identifier that cannot be changed over time. A URN cannot be assigned to other resources even if the mapped resource has ceased to exist. RFC 3188 [10] describes URN:NBN.

² Other permanent identifiers are DOI and Handle.

The domain urn:nbn:se is managed by the Royal Library. The Royal Library assigns URNs in the Swedish national bibliographic number domain (URN:NBN:se) to organisations and the public in Sweden [6].

The DiVA archive has been assigned the sub domains URN:NBN:se:X:diva where X is an acronym of the participant in the project. Uppsala University is represented by acronym uu. The sub domains are managed locally. To automatically assign every published document in the DiVA Archive an unique URN:NBN identifier is a locally created and managed serial number added to the sub domain, e.g. URN:NBN:uu:se:diva-3100.

Implementation of an URN:NBN Resolution Service

The core purpose of URN:NBN is to identify and locate the resource on the Internet. To be able to use the URN:NBN as an identifier on the Internet, a resolution service is needed. To resolve an URN:NBN within the Swedish domain, a resolution service exists at the Royal Library. The current implementation is a prototype with only basic functionality: to resolve URN:NBNs and redirect the requests to an URLs.

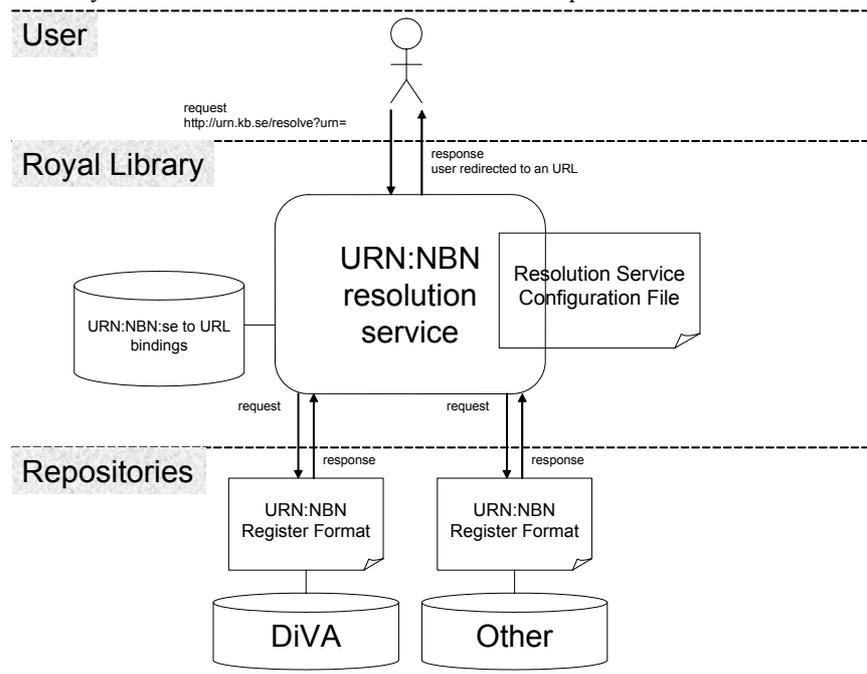


Fig. 3. Architecture of the URN:NBN implementation

The resolution service is implemented as a java servlet and contains a harvester, which can harvest URN-URL-bindings from many different repositories, each defined by an URI. The servlet needs a configuration file with information about where

repositories are located and also a time value for each repository, which determines the time between two harvestings. The URN:NBNs in the repositories are represented in an well defined XML format, which are stored as a file either at the Royal Library or at the institution where the archive is located. The file stored at the Royal Library contains individual assigned NBNs [6]. With this model the administrative part of the creation of URN:NBNs can be easily distributed to many different archives and can be automated without the need of human interaction.

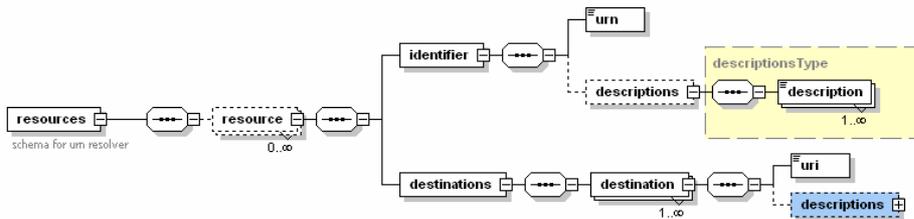


Fig. 4. XML schema defining URN:NBN Register Format used for the Resolution Service

The Role of URN:NBN in the DiVA Archive

The URN:NBN has several roles in the DiVA Archive and the archiving workflow. We very quickly recognized that it would be practical to use the same identifier, if possible, within the entire system. The main role of URN:NBN is of course still to identify and locate the resource – to give a stable point of reference for the documents, but in addition to that the URN:NBN is also used as an unique identifier within the archive. The URN:NBN is currently also used as a part of naming convention for files and directories in the archive and the archival packages, that are stored in the local long-term archive and transmitted to the Royal Library [8].

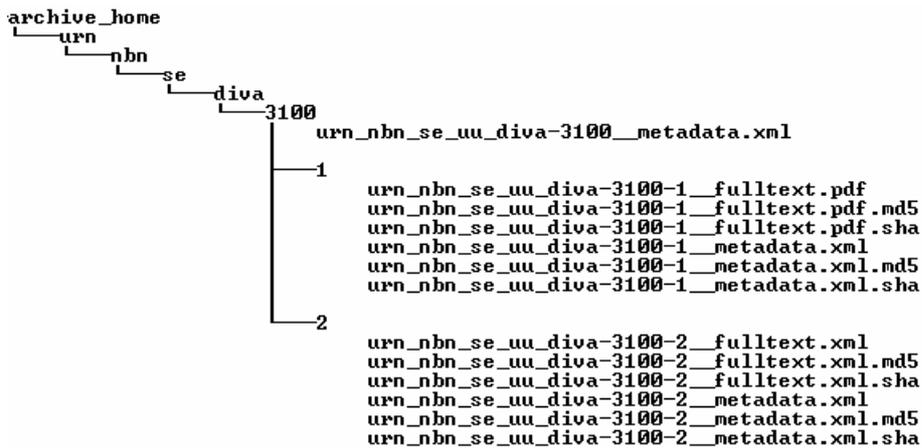


Fig. 5. URN:NBN is used as a part of the naming convention for files and directories in the archive

DiVA uses URN:NBN as a primary identifier for each item. An item is defined as a single publication without consideration of its format (i.e., both the electronic and printed versions of a single thesis can be referenced on the metadata level by the same URN:NBN). Currently this is important because of 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.

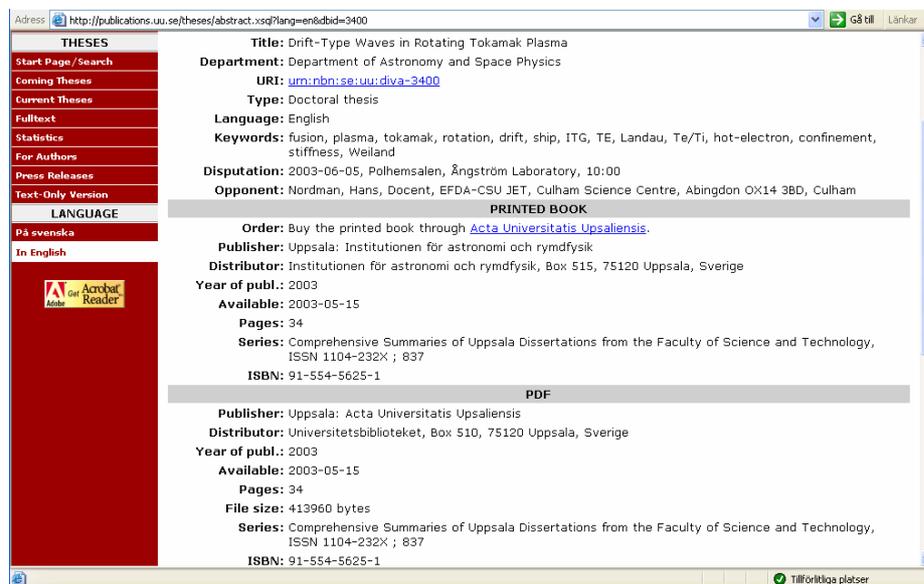


Fig. 6. Example of a user interface for an item with different manifestations

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.

As mentioned earlier, the resolution service has only a basic functionality, but there are plans for further development. A network of URN-servers, like the network of

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

Archiving Packages and the Workflow to the National Library Archive

In the DiVA Archive, all administrative and descriptive metadata are stored in XML that conforms to the DiVA Document Format [3]. 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). When the document is published, checksums for the files in the related manifestation are built and the metadata are locked and stored in the folder of the corresponding manifestation.

Based on experience, changes in the metadata after the date of publishing of document occur. For this reason documents are not sending at the publishing date to the long-term preservation archive immediately. After a certain time (currently 2 months) from the publication date, each manifestation will be 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 will be check summed and stored in the local archive for long-term preservation.

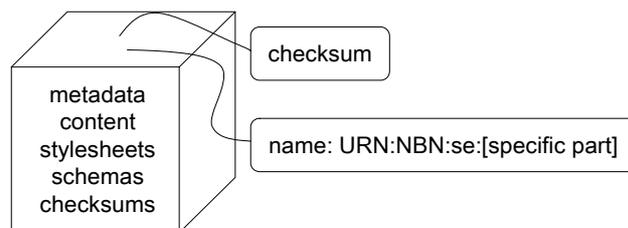


Fig. 7. Archival package

Technical solutions supporting this part of the workflow are currently under development and the workflow will be fully implemented during the autumn this year. Packages and checksums will be transmitted 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 reduce the risk of data loss.

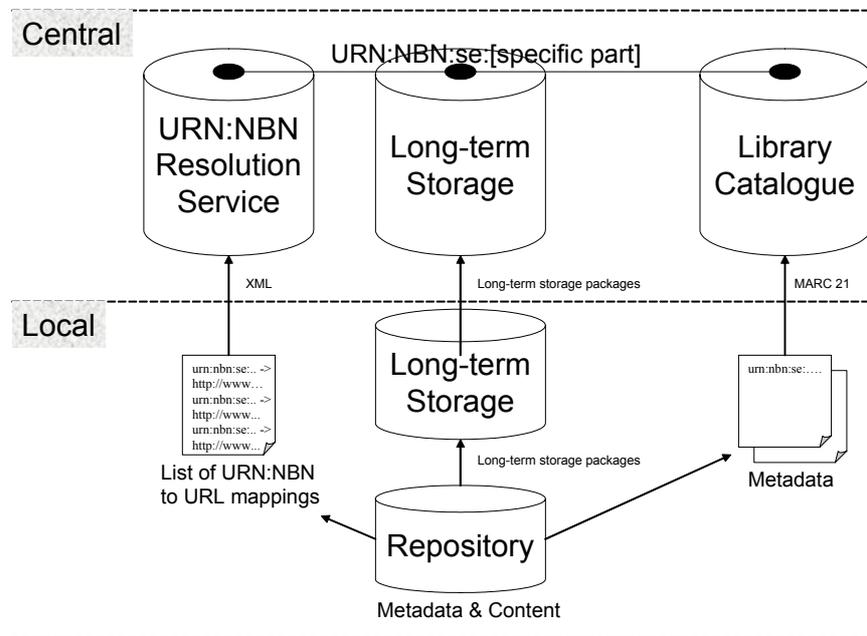


Fig. 8. Archiving workflow between the local repository and the National Library

Related Work and DiVA Project Contribution

The DiVA Archive and its associated workflow build on previous research and development in areas of digital archiving and digital library systems. The cooperation with the Royal Library and experiences of the library community with URN:NBN and systems for delivery of the archival depository copy of electronic documents are the knowledge base we build on. The Reference model for an Open Archival Information System (OAIS) [2] and the results of the project CEDARS [1] are the work we found most useful in the context of designing the workflow and building the DiVA long-term preservation archive.

Our team has applied these and other concepts in the development of a low-cost system that supports a fully automated workflow (from point of submission). Using the harvesting model for updates to the mapping registry for the resolution service makes the management of URN:NBN simple. Automatic creation of the MARC 21 record (including URN:NBN) makes cataloguing faster and less expensive. Using the “push” model to deliver archival packages to the National Library Archive makes this process more reliable and easier to manage. The use of XML for all metadata associated with each archival package increases the likelihood of future understanding

of digital objects in the archive and offers the potential to easily extract document metadata, if necessary.

Because the technical solutions underpinning the system are built using component-based design methodology, modules can be seamlessly replaced with improved components. The resulting modularity and component reusability offers a tremendous advantage over other methods and is a solid basis for further local and cooperative development.

Conclusions and further work

The archiving workflow and the DiVA Archive presented in this paper are examples of practical solutions that help demonstrate that long-term access to the institutional research publications can be assured with cooperation from national libraries.

Our experience with this workflow is now the basis for a new project funded by the Royal Library's department for National Co-ordination 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 - for example: What is a minimal level of preservation metadata - will be explored. The project will also look into some other standards (e.g., METS [7]) as a possibility for package description. Because of a lack of practical examples of implementations within the library community, we believe this project will be broadly useful.

References

1. CEDARS web site - <http://www.leeds.ac.uk/cedars/>
2. Consultative Committee for Space Data Systems, Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-B-1, Blue Book. January 2002. <http://www.ccsds.org/documents/650x0b1.pdf>
3. DiVA Document Format – XML schema: <http://publications.uu.se/schema/1.0/diva.xsd>
4. Electronic Publishing Centre at Uppsala University Library – web site <http://publications.uu.se/epcentre/>
5. Kungliga Biblioteket – Swedish National Library – web site <http://www.kb.se/>
6. Kungliga Biblioteket – URN:NBN – <http://www.kb.se/urn/>
7. Metadata Encoding and Transmission Standard (METS). <http://www.loc.gov/standards/mets/>
8. Müller, E., Klosa, U., Hansson, P., Andersson, S., Siira, E.: Using XML for long-term Preservation. Experiences from the DiVA project. ETD 2003 in Berlin. <http://publications.uu.se/etd2003/papers/LongTermPreservation.pdf>
9. Müller, E., Andersson, S., Klosa, U., Hansson, P.: Metadata Workflow Based on Reuse of Original Data. ETD 2003 in Berlin. <http://publications.uu.se/etd2003/papers/MetadataWorkflow.pdf>

10. RFC 3188 . Using National Bibliography Numbers as URNs .October 2001
<http://www.ietf.org/rfc/rfc3188.txt>
11. Uppsala University Publications – <http://publications.uu.se/>