Metadata Workflow Based on Reuse of Original Data

Eva Müller
Stefan Andersson
Uwe Klosa
Peter Hansson
Electronic Publishing Centre, Uppsala University Library
epub@ub.uu.se
Uppsala University Library, Box 510, SE-751 20 Uppsala, Sweden
publications.uu.se/
Keywords: Metadata, XML Schema, Library catalogues, MARC 21

Abstract

The goal of the DiVA project is to develop a workflow where information from the original document created by the author can be reused to extract metadata for various purposes.
There are two ways to submit an item and store the metadata in the SYSTEM:
1. using a template (MS Word, Open Office, Star Office, LaTeX)
2. updating through the DiVA Manager application

In both cases metadata are converted to XML in the DiVA Document Format and stored in the DiVA SYSTEM.
Metadata can be generated from the DiVA Document Format in many different formats: currently in Dublin Core (HTML-text and XML-RDF), TEI-Header, Endnote and Reference Manager, and MARC 21. All records are also presented in unqualified Dublin Core and MARC XML to support harvesting by OAI-PMH service providers.
A workflow has been created where the information, originally created by the author, forms the basis for the bibliographical description of the record in the Swedish National Library SYSTEM as well as in the local library SYSTEM at Uppsala University Library.
The record is transmitted in MARC XML format to the Royal Library in Stockholm and converted to MARC 21 "tape format" for import into LIBRIS, the national library SYSTEM and, after adding local information such as holdings, exported to the local SYSTEM. Records can also be downloaded directly from the DiVA web site in MARC 21 "tape format" (ISO 2709), as well as MARC XML format, for use in local library SYSTEMs.
Each item in the DiVA archive is assigned a unique, persistent Uniform Resource Name and National Bibliographic Number.
In addition to the delivery of metadata to the National Library, full-text files (currently in PDF together with a DiVA Document Format file containing all archiving metadata) will also be delivered to support long-term preservation.

Preface

DiVA-Digitala vetenskapliga arkivet (DiVA Archive)-is a comprehensive description of a searchable archive containing all documents which are published in an electronic form at Uppsala University in Sweden. Other Swedish universities are also co-operating in the project within the DiVA framework. One part of this archive is the database containing theses published at Uppsala University from 1998 to date.

In September 2000 an Electronic Publishing Centre was established at Uppsala University Library. Its primary assignment was a project in which technical solutions, and a well-functioning workflow, for electronic posting and full-text publication of doctoral theses, essays, working papers and other types of scientific publications were to be created.

The first phase of the project was completed in 2002 and the result was the DiVA Publishing System - a SYSTEM for electronic publishing of different types of publications.

Metadata records from the DiVA SYSTEM are distributed to services which are relevant for the dissemination of information about the research activities of Uppsala University and other participating partners.

This paper will describe the creation process in general of metadata regarding doctoral theses in the DiVA SYSTEM, and more specifically how the data submitted by the author can be reused for bibliographic information in library SYSTEMs.

The idea of reusing of information from the source publication is not completely new. We can see a connection to CIP[1] (Cataloguing in Publication - an old idea from the seventies), where a catalogue record, which could be reused for the cataloguing of the actual item, was printed directly on the verso of the title page of the source publication. Although the records could be distributed in machine readable form, the initial cataloguing was always carried out by a cataloguer and quite often the information would be retyped by another cataloguer on a card or in a library SYSTEM. This was a part of our inspiration when we started thinking of a new workflow for electronic publishing and about the possibilities of reusing of the information submitted by the author.

Three years ago, when we started experimenting with this new workflow, cataloguers were not completely happy about the idea, at least not in the case of Uppsala University Library. Some of them were very sceptical to the capabilities of the new technology to produce good quality records, some of them were sceptical to the skills of our developer team; some of them were simply wor-

1 See: http://cip.loc.gov/cip/

ried about the future of the cataloguers as a profession. However, the attitude has changed in the course of time. The DiVA development team consists not only of SYSTEM developers and graphic arts designers, but also of former cataloguers. This made it easier to communicate within the library. Now this workflow is highly accepted and our colleagues can concentrate on more qualified tasks when cataloguing theses (indexing and classification) instead of retyping the bibliographical information.

The results presented here have been achieved in cooperation with the Royal Library - the National Library of Sweden.

Submission and Creation of Metadata for Theses in the DiVA System

Two kinds of doctoral theses are produced at Swedish universities: monographs and summaries. The summaries consist of a number of individual papers and an independent summary. In Uppsala the summaries are primarily published in five series, one from each faculty, called "Comprehensive summaries of Uppsala dissertations from the faculty of …"

The Vice-Chancellor of Uppsala University has through decisions in 2000 and 2002 stipulated that all new theses must be posted electronically and that all comprehensive summaries are to be published on the Internet. Monographs can be published electronically on a voluntary basis.

Posting

Posting on the Internet of doctoral theses at Uppsala University must take place at least three weeks before the day when the thesis is to be publicly defended, and the thesis must also be available in printed form for reviewing at the university library.[2]

The basis for posting can be delivered as a file created from a template (developed for the DiVA Publishing System) *or* as another file that includes the original text (if the doctoral student does not have access to any of the platforms used for the templates or cannot use the templates at all). In both cases the file must be sent to the Electronic Publishing Centre at least a couple of days in advance, in order for the doctoral student to obtain a receipt when the printed copies of the thesis are delivered at the University Library, which is a condition for posting.

Templates

Form-based templates for MS Word, Open Office, Star Office, and LaTeX have been developed for posting. To assure high quality metadata several controls are added to the forms including drop-down menus with controlled values for names of publication types, degrees, departments, addresses, series, and distributors. Other templates are used for the creation of the actual contents (the full-text document).

The templates used for posting produce XML files that contain all metadata regarding the thesis, the author, and the public defence of the thesis. The XML files are stored in the DiVA Document Format. Through XSLT and XSL-FO these files are used in various contexts where metadata needs to be extracted, e.g. for the creation of title pages, cover pages, edition notices of the printed and electronic publications; web pages and so on. (See Fig. 1).

The DiVA Document Format

The DiVA Document Format is an internal metadata format that was developed in the project since other existing formats considered did not include all the features needed for the DiVA Publishing System. The format, described in XML Schema[3], is component based and extensible. Some inspiration has been gathered from the work concerning Functional Requirements for Bibliographic Records, FRBR[4], by IFLA. For instance all formats of the document (printed as well as electronic ones) are described within the same record as "manifestations".

From the DiVA Document Format also the necessary bibliographic data for library SYSTEMs, which will be closer looked at below, can be created.

The Cataloguing Process at Swedish Research Libraries

LIBRIS - the Union Catalogue of Swedish Libraries

Research libraries in Sweden primarily catalogue their resources in LIBRIS - the union catalogue of Swedish libraries - but use local library management SYSTEMs for circulation and patron information. This means that the bibliographical records for books, periodicals, electronic documents, and other publications are registered only once in the union catalogue. The participating libraries will then add local information on their specific holdings, locations and subject headings. Afterwards the records are exported to the local library SYSTEM where item-specific data, e.g. a barcode, is appended. At the moment

2 See: http://info.uu.se/Internt.nsf/5c6cb794ca96404fc125680e004836ad/56588f26f92ec1d0c1256c6e00456886/$FILE/Beslut.pdf
3 See: http://publications.uu.se/schema/1.0/diva.xsd
4 See: http://www.ifla.org/VII/s13/frbr/frbr.htm

some 200 university and special libraries are using LIBRIS for cataloguing and more than 1,400 libraries are using it for interlibrary lending. The union catalogue, comprising 4,5 million titles, is also publicly available on the web. The LIBRIS-department at the Royal Library in Stockholm is responsible for the administration and development of the SYSTEM.[5]

MARC 21

Since January 2002 the Voyager software from Endeavor Information Systems is utilised as the library management SYSTEM for LIBRIS, replacing an older SYSTEM. In connection with this change of SYSTEMs it was also decided to replace the local LIBRISMARC/LIBRIS III format with MARC 21 (the harmonised USMARC and CAN/MARC formats published in a single edition by the Library of Congress and the National Library of Canada in 1999).[6] Facilitating the exchange of bibliographical information on an international level was one of the reasons behind this decision.[7]

Cataloguing Swedish Theses

Normally the theses published at Swedish universities are initially catalogued in LIBRIS by the local university library. The university libraries generally use the minimal level (code 7 in MARC Leader) for the bibliographic information. In KRS[8], the Swedish cataloguing rules, which are a translation and revision of the Anglo-American cataloguing rules, second edition, this level is described in §1.0D2.

Traditionally theses (and other university publications) are catalogued in the same way as other items acquired by the university libraries: A cataloguer will receive the book, search for it in the LIBRIS database, and, if it is not already registered, type in the full bibliographical record including all ISBD(G)-punctuation (General International Standard Bibliographic Description).[9]

Subsequently the theses will also be included in Svensk bokförteckning[10], the part of the Swedish national bibliography of literature which covers monographs issued in Sweden. The bibliography is based on copies supplied to the Royal Library by publishers and other organisations and on legal deposit copies. It is produced from the LIBRIS database. The bibliographic information will, in connection with this, be refined by the section of National Bibliography Monographs at the Royal Library to meet the full level as described in §1.0D3 of KRS. As an example it can be mentioned that the number of Swedish theses issued in a single year (2002) and registered in the Swedish national bibliography was 2,247.

New DiVA Workflow

The main intention of the new cataloguing workflow implemented through DiVA is naturally to reuse the original data which was created by the author (as described above, 1.1-1.2) as the basis for the bibliographical record instead of typing the same information all over again.

Additionally, the bibliographic record will be available in the national union catalogue more or less simultaneously as the thesis is made publicly available. Otherwise a small sample of Uppsala theses showed that it will typically take at least a week before the thesis turn up in LIBRIS (if cataloguing is done manually).

In the complete DiVA workflow described above where the metadata entered by the author actually forms the title page and edition notice of the publication (in print or on the Internet) one can also be sure that the bibliographic information is the correct one - at least from a cataloguing point of view! - since it is reproduced directly from the source document.

Creation of MARC 21 Records in DiVA

The records created from the information submitted by authors, and stored as files in the DiVA Document Format as described above, constitute the starting point for the creation of the MARC 21 records in the DiVA Publishing System. Transferring the records directly from DiVA to LIBRIS means that cataloguers do not have to import them from another source or use any other software than the Voyager cataloguing client to find them. (See Fig. 2).

At first an XSLT stylesheet will transform the record in the DiVA format into another XML file in the MARC XML format, the new XML exchange format for MARC as published by the Library of Congress.[11] The XML format is used because the OAI proctocol for metadata harvesting (OAI-PMH)[12] is used to collect and deliver the records which are to be sent to LIBRIS. Besides, no further scripting, apart from the XSLT transformation, is needed to create a full MARC record. A MARC ISO-2709 record can then be created without data loss from the MARC XML record. Leader data positions not needed in the XML environment are retained as place holders containing the value 0. The MARC data fields are created by matching DiVA elements to MARC elements. Standard phrases are included in the XSL templates where ap-

5 See: http://info.libris.kb.se/infosvensk/allmaninfo.htm
6 See: http://lcweb.loc.gov/marc/
7 See: http://info.libris.kb.se/infosvensk/Nyheter_Fakta/librisnytt/librisnytt34.htm#rubrik1
8 Katalogiseringsregler för svenska bibliotek. Lund 1990
9 See: http://www.ifla.org/VII/s13/pubs/isbdg.htm
10 See: http://www.kb.se/nbm/svb.htm
11 See: http://www.loc.gov/standards/marcxml
12 See: http://www.openarchives.org/OAI/openarchivesprotocol.html

propriate. Information that appears more than once in the MARC record can be created from one source (e.g. both data fields 100a and 245c can be created from a single DiVA creator element).

Since it has been decided to store all the ISBD(G)-punctuation (even the punctuation between fields) in the actual fields of the LIBRIS records, rather than letting the user interface display them, the punctuation is created in the MARC XML record through a series of conditional tests in the XSLT stylesheet, otherwise it must be added manually by the cataloguer. Quite often a number of conditions must be tested.

The date of birth is required for Swedish citizens in main and added entry fields (100/700) for personal names in LIBRIS. As the date is submitted by the authors themselves the cataloguer will not have to look it up in the university directory.

The current version (1.0) of the DiVA Document Format supports all corresponding MARC 21 fields and indicators for the bibliographic description except the number of non-filing characters[13] which occurs as the second indicator in the title statements, e.g. fields 245 and 440. Because there is no way of describing alternative filings of titles in the DiVA SYSTEM at present, the second indicator will always initially be set to 0 and may have to changed manually later on.

Transferring DiVA Records to Libris

The OAI Harvester application from OCLC[14] contacts DiVA daily to ask for newly published theses. As a response to the request the DiVA SYSTEM delivers metadata records conforming to the MARC XML format described above. The records are stored in an XML file which is delivered to LIBRIS by a file transfer protocol over the Internet.

The MARC XML records will eventually be converted to MARC ISO-2709 ("tape format"), and imported into the LIBRIS database. The conversion will calculate the numeric strings for the record length (Leader position 00-04) and Base address of data (Leader position 12-16) and replace the 0-values of the MARC XML record. It also establishes the directory of the ISO 2709 record. Since the Voyager library management SYSTEM does not support Unicode the character encoding must also be converted to ANSEL. These conversions can also easily be accomplished by the MARC4j API[15].

Persistent Links Through URN:NBN

An important feature of the DiVA Publishing System is the URN:NBN (Uniform Resource Name National Bibliographic Number) which is created in co-operation with the Royal Library. Each item in the archive is assigned a unique, persistent National Bibliographic Number and Uniform Resource Name to assure future access.

Two 856 fields for electronic location and access are added in MARC 21 records: one for the URN:NBN and one for the corresponding URL to the resolution service. The latter one can be used for automatic creation of web links in library SYSTEMs.

Downloading of Individual Records in Different Formats

Metadata of individual theses can be downloaded from the DiVA web site in many different formats, such as Dublin Core[16] (HTML-text[17] and XML-RDF[18]), TEI-Header[19] and specific formats for bibliography software like Endnote[20] and Reference Manager[21]. Records in MARC 21 can also be downloaded for use in any local library SYSTEM. If the publication is available in an electronic form a separate record for the electronic document can be created. The bibliographic records are available in three different formats: MARC ISO-2709, MARC XML, and MODS[22]. Authority records for the authors can also be downloaded.

All records are also expressed in unqualified Dublin Core and MARC XML to support harvesting by OAI-PMH service providers[23]. Selective harvesting through sets of subjects, universities, and document types is also available.

13 A value that specifies the number of character positions associated with a definite or indefinite article (e.g., Le, An) at the beginning of a title that are disregarded in sorting and filing processes.
14 See: http://www.oclc.org/research/software/oai/harvester.shtm
15 MARC4J is an easy to use Application Programming Interface (API) for working with MARC records in Java. The API consists of an event-based MARC parser, an object model for in-memory editing of MARC record objects and SAX2 based producers and consumers for conversions between MARC and MARC XML. Available from http://marc4j.tigris.org/
16 See: http://dublincore.org/
17 See: http://www.ietf.org/rfc/rfc2731.txt
18 See: http://dublincore.org/documents/2002/07/31/dcmes-xml/
19 See: http://www.tei-c.org/P4X/HD.html
20 See: http://www.endnote.com/
21 See: http://www.refman.com/
22 See: http://www.loc.gov/standards/mods/
23 See: http://www.openarchives.org/service/listproviders.html

Archiving

The Swedish Legal Deposit Act[24] does not cover online documents. At present the possibility is being examined how to collect those documents with Swedish connections. In the meantime, an agreement[25] was set up in September 2001 between the Royal Library and Uppsala University where the long-term preservation of electronic publications from Uppsala University is guaranteed. Regarding doctoral theses this is particularly important if the printed edition **falls short of around 30 copies, which is the current limit for legal deposits used by the Royal Library** as far as printed documents are concerned. The full-text files from the DiVA Archive (currently in PDF) are sent to the Royal Library for archiving together with the full DiVA Document Format file containing all metadata.

Conclusions

At Uppsala University alone about 450 doctoral theses, thousands undergraduate dissertations, working papers and other publications are produced yearly. An increasing number of them are now published electronically.

A great advantage of the DiVA Publishing System is that it is scalable and can handle the whole workflow on a large scale. That means that high quality metadata can be produced directly from the templates even for undergraduate dissertations and other low priority material. This allows a higher level of resource discovery of this type of material. In the traditional library workflow, when cataloguing is done completely manually, these types of documents - non printed documents - would probably not be catalogued at all.

Appendix

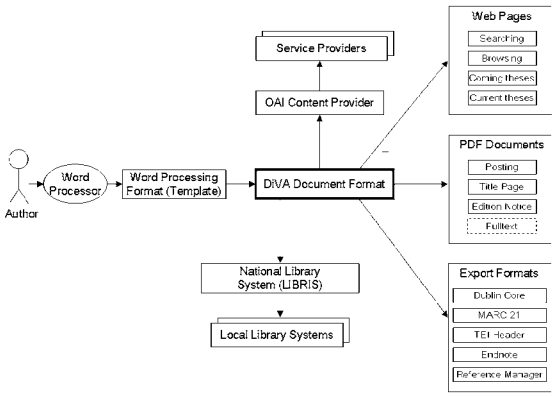

Fig. 1: Metadata from the author used in many different contexts.

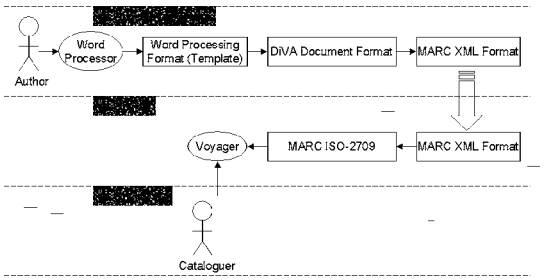

Fig. 2: Metadata flow from the author to the cataloguer.

24 See: http://www.kb.se/ple/sfs.htm
25 See: http://www.ub.uu.se/diverse/avtal.pdf