The DiVA System: Current Status and Ongoing Development

Uwe KLOSA

Electronic Publishing Centre, Uppsala University, Sweden uwe.klosa@ub.uu.se

Eva MÜLLER Electronic Publishing Centre, Uppsala University, Sweden eva.muller@ub.uu.se

INFORUM 2006: 12th Conference on Professional Information Resources Prague, May 23-25, 2006

Abstract: The DiVA system, originally developed at Uppsala University, has evolved out of a scholarly repository and publishing system solely used by one university into a system used and further developed in collaboration with seventeen universities in Sweden, Norway and Denmark. The system consists of several modules which are entirely built on XML and Java technologies. This modularisation facilitates the possibility of enhancement of the system. Moreover, it increases the scalability of the system.

This paper describes the current status of some of these modules. Furthermore it gives an insight into the ongoing development.

The DiVA publishing system supports work flows for printing as well as digital publishing of scholarly documents. The repository holds metadata and full-text documents for doctoral theses, master and student theses, research reports, and articles. The metadata and files can be accessed freely via a joint portal, local search services, and with OAI-PMH. The metadata and files are packaged and sent to the National Library of Sweden in Stockholm for archiving and also archived locally in a dark archive.

The DiVA consortium is at present engaged in a discourse on the usage of DiVA as a repository for post- and pre-prints. In addition to that, the development team researches for ways to evolve DiVA from a dark archive into an archive implementing the OAIS model.

Introduction and Background

In 2000, Uppsala University¹ decided to start a project whose main objective would be to publish all scholarly work of the university electronically. The *Electronic Publishing Centre* (henceforth referred to as EPC) has been operational since September 2000. The main goal of the EPC is to support researchers and students of the university with user-friendly and effective tools and to help them publish their research results, as well as to create solutions for publishing and archiving. This institutional repository contains digital documents published at Uppsala University which can be searched in a structured form simultaneously with the library catalogue and other relevant search services and to offer metadata to information services which are relevant to the dissemination of information regarding research in Uppsala".² This was extended to also provide the access, use and understanding of digital objects in the long term.

It was decided to build a system based on open standards, and to incorporate open source software where possible. Moreover, we opted for XML³ and Unicode⁴ for the metadata files and for the full-text files as well. As we have always been aware that it would take some time to accomplish these goals, we decided to use PDF⁵ for the full-text files initially and to implement XML support from the very beginning.

We developed the DiVA system. DiVA is the acronym for *Digitala Vetenskapliga Arkivet* which translates as *Academic Archive Online*. DiVA has evolved out of a scholarly repository,

¹Home page Uppsala University. <<u>http://info.uu.se/fakta.nsf/sidor/uppsala.university.id5D.html</u>>

²Home page Electronic Publishing Centre. <<u>http://epc.ub.uu.se/#english</u>>

³XML specifications. <<u>http://www.w3.org/XML/</u>>

⁴Unicode home page. <<u>http://www.unicode.org/</u>>

PDF reference. <<u>http://partners.adobe.com/public/developer/pdf/index_reference.html</u>>

publishing system solely used by one university, into a system used and further developed in collaboration with seventeen universities in Sweden, Norway and Denmark. The system consists of several modules which are entirely built on Java and XML technologies. This modularisation facilitates the possibility of enhancement of the system by replacing modules with new enhanced modules. Moreover, it increases the scalability of the system.

In 2002, the universities using DiVA formed a consortium for the purpose of sharing solutions, tools, expertise and costs. This cooperation is open to all universities and to publicly financed research departments.

The DiVA System

After it had been operational for testing and further development at Uppsala University for a year and a half, the DiVA system became fully operational in January 2003. The system consists of several modules: a repository, a publishing system, authoring tools, a joint portal⁶, search services, dissemination services, and a dark archive. Furthermore, modules supporting pre-publishing work flows as well as archiving work flows are under development.

The DiVA repository holds metadata and full-text documents for doctoral theses, master and student theses, research reports, and articles. The repository stores the files directly in the file system; it is installed on an independent server and can be accessed remotely by all other modules. The access to this remote interface is restricted with the help of a cryptographic key pair. Only modules which send a key recognized by the repository get access.

The DiVA publishing system supports work flows for printing as well as electronic publishing of scholarly documents. There are various tools for metadata submission, such as a web interface where authors can submit their documents and the metadata which go with them. Metadata can also be submitted with the help of a template for MS Word, and can be imported and updated via an administrative interface. The metadata are reused for the creation of title pages, book covers and web pages.⁷ Furthermore, subsequent to the publication of a document, the metadata are automatically sent to the National Catalogue⁸ at the National Library of Sweden⁹, Stockholm (see figure 1).¹⁰

The part which will handle the publishing of full-text files in XML is still under development. Thus far it is possible to create XML files by XSL transformation of documents formated by templates for various word processors. For our purposes we developed a work flow where we use Open Office¹¹ for converting incoming documents to the Open Document Format¹², and then a customised XSL style sheet transforms this into an XML format derived from DocBook¹³ containing the entire document. These DocBook documents form the basis from which the table of contents and the references are published. If the entire document were to be published from XML, the authors would have to check the text and to authorise the publication once again. The authorisation work flow is still under development.

The metadata and full-text files can be accessed freely via a common interface, the DiVA portal. As DiVA supports the Open Archives Protocol (OAI-PMH¹⁴) and RSS¹⁵, the metadata can be accessed and reused in many ways. Furthermore, there are some predefined search queries implemented, such as a list of coming theses and a list of current publications. It is possible for local departments and even individuals to include search results on their own home pages, either by including predefined html snippets or with RSS.

⁶DiVA portal. <u><http://www.diva-portal.org/index.xsql?lang=en></u>

⁷Klosa, U. Siira, E. Andersson, S. (2005)

⁸Libris home page. <<u>http://www.libris.kb.se/english/indexeng.jsp</u>>

⁹Home page National Library of Sweden. <<u>http://www.kb.se/ENG/kbstart.htm</u>>

¹⁰Andersson, S. Hansson, P., Klosa, U. Müller, E. (2003) ¹¹Open Office home page. <<u>http://www.openoffice.org/</u>>

¹²Open Document Format specifications. <<u>http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=office</u>> ¹³DocBook home page. <<u>http://www.docbook.org/</u>>

¹⁴Open Archive Initiative Protocol for Metadata Harvesting. <<u>http://www.openarchives.org/documents/index.html</u> > ¹⁵Really Simple Syndication (RSS). RSS 2.0 Specification. <<u>http://blogs.law.harvard.edu/tech/rss</u>>.

Each university using DiVA has its own local instance of the DiVA system and its interface adapted to its own graphic profile. Those local services contain in many cases also metadata for documents which are not published in full-text.

The initial version of the DiVA system used an Oracle database for storage, indexing and searching parallel with the storage of metadata in XML files in the file system. However, in course of recent work, this Oracle based search module has been replaced with a new search module which employs the Apache Lucene search engine¹⁶, an open source project.

DiVA Document Format

Within the DiVA project we developed a document model for describing scholarly publications. The model is described by the *DiVA Document Format* (DDF) which is expressed in an XML schema. This schema is fine grained, hierarchical, modular and extensible. It combines metadata elements with elements for structural mark-up and makes it possible to store metadata with a high level of granularity. When developing the format, we drew inspiration from the work concerning Functional Requirements for Bibliographic Records (FBFR)¹⁷. An application profile describing the format has been published recently, which provides deep documentation and details on mappings to other formats.¹⁸

Two important requirements for the format have been the compatibility with other metadata formats and standards and that it be human-readable.

The current version of the DDF contains more than 100 elements. Some elements are containers which embed one or more elements of the same type (the creators container, for instance, embeds one ore more creator elements). Most of these embedded elements can contain an identifier which could be for example an identifier from an authority system, such as a university directory containing names of departments, units and employees.

DDF is used for the storage of metadata and full-text files in XML, inter-application communication, the creation of web pages, and as a basis for other metadata formats. The metadata can be reused as they are or transformed into a standard format, such as Dublin Core, METS, Marc and MODS. Furthermore, the transformation into import formats of reference managing tools is also supported.

Access in the Long Term

As stated above, one of the goals of the DiVA project is to ensure the access, use and understanding of digital objects in the long term. A prerequisite for this is to have a persistent unique identifier for each resource. In view of the fact that we collaborate with the National Library of Sweden in Stockholm, we decided on URN:NBN for this purpose. An identifier of this type is useless without a service that knows which identifier is linked to which resource. Therefore a resolution service has been created in cooperation with the National Library of Sweden. From each DiVA repository, as well as other, non-DiVA repositories, the resolution service harvests the mappings between URN:NBN and URL (see figure 2). The current version of the resolution service provides also the resolution of other identifier schemas, such as DOI¹⁹, Handle²⁰ and ARK²¹ identifiers, which is done by a redirect to relevant services.

The URN:NBN is not only used as an identifier for the object, but is also used as the name convention for metadata files, as well as full-text files and their attachments.²² Even the archiving package which is sent to the National Library of Sweden for long term preservation contains the URN:NBN in its name. Each package consists of the metadata files, full-text files, and attach-

¹⁹DOI. <<u>http://www.doi.org</u>>

¹⁶Lucene home page: <<u>http://lucene.apache.org/java/docs/</u>>

¹⁷FBFR. <<u>http://www.ifla.org/VII/s13/frbr/frbr.htm</u>> ¹⁸DiVA application profile. <<u>http://epc.ub.uu.se/diva-app-profile/</u>>

²⁰Handle. <<u>http://www.handle.net/</u>>

²¹ARK. <<u>http://www.cdlib.org/inside/diglib/ark/</u>>

²²Hansson, P. Andersson, S. Klosa, U. Müller, E. Siira, E. (2003)

ments of one resource. At present the metadata in this package are not enriched with administrative or format specific metadata.

Collaboration

The DiVA system was originally developed solely for Uppsala University. When other universities showed their interest in DiVA, however, a consortium was formed which consisted of five Swedish universities in the beginning. The consortium is open to all universities and publicly financed research departments. In 2003, the consortium was joined by two universities from Denmark and Norway, and today consists of seventeen universities altogether (see table 1).

To ensure the consortium's efficiency and sustainability clear terms of participation and good communication have been adopted.²³ The participation is based on a written agreement between each University and Uppsala University. Conditions for participation and cost regulation are laid down by this agreement. The fee which every participant is charged when joining the consortium flows directly into the development budget. Each participant has the choice to either operate their DiVA system locally or to use a hosted system at Uppsala University which is the option chosen by most participants.

Good communication between all participants is maintained by the following means:

- A helpdesk ensures that the system administrators at the participating universities receive all support they might need.
- A Wiki where the participants find system documentation and FAQ. Here they can also add their ideas or anything which might be of interest for other participants.
- Each system administrator gets an introduction into the functionality of the system.
- Twice a year user group meetings are held where working groups are formed, for example, to create specifications of requirements for further development.
- A mailing list affords an open channel for discussions concerning electronic publishing in general.
- A newsletter which informs about the latest development of the system and coming conferences, as well as any news concerning DiVA.
- An area where system developers who are involved in the development of DiVA, and those who want to develop and share own modules for the DiVA system, can meet and exchange ideas and knowledge.

Apart from the collaboration within the consortium there is also cooperation with partners outside the DiVA community. The cooperation with the National Library of Sweden, for instance, resulted in an agreement on the archiving of documents stored in DiVA and the creation of a resolution service for URN:NBN.

Other types of cooperation often result in research projects which are related to the area of electronic publishing, long-term preservation and content markup in general. Theses projects are often on a national level. International cooperation thus far has resulted in the exchange of knowledge and ideas.

Future and Ongoing Development

Ever since DiVA became operational, we have been gathering specifications of requirements for the next generation of DiVA. For about half a year, the development of new modules has been in full swing.

The recent launch of the new search and indexing module made us independent of the Oracle database which we used in the very first version of the DiVA system. Even though we still store authority records in a relational database, we can now choose between a variety of different

²³Müller, E. Sandgren, F. Andersson, S. Klosa, U. Hansson, P. (2005)

database management systems. The new search and indexing module employs Apache Lucene. It automatically indexes all metadata and full-text files.

In due course, the XML representation of the DiVA Document Format will be replaced with a new format based on a complex object format, such as DIDL²⁴ or METS²⁵. The current XML schema has a high level of granularity, but it lacks the structural flexibility of such a complex object format. While in DIDL, for example, one can have elements of the same type at different levels in the hierarchy, this is not possible in the current version of DDF.

Another reason for the creation of a new XML schema is that the DiVA consortium is at present engaged in a discourse on the usage of DiVA as a repository for post- and pre-prints. Furthermore, the development team is researching for ways to evolve DiVA from a dark archive into an archive implementing the OAIS model. The structure of the complex object format will facilitate the addition of other types of metadata, such as rights metadata, administrative metadata, and file format specific metadata collected from a format registry, such as Pronom²⁶ or the Global Digital Format Registry (GDFR)²⁷.

In addition, we are investigating the possibility of an integration of a more flexible repository solution for digital objects, such as for example the Fedora repository²⁸. Such a repository would give us the opportunity to define multiple views on each digital object stored in the repository. An image, for example, could then be stored in a single format and the definitions of transformations to other formats could be stored with this image. Images in these formats would not have to be stored in the repository. In the current version of DiVA, this functionality is part of the user interface. Changes in the metadata format or file format always affect the web interface and administrative interface directly and, consequently, the development of different modules has to be constantly synchronised. The usage or implementation of a repository like Fedora would facilitate the separation of user interface and repository.

In parallel with the work described above, techniques to publish both electronically and traditionally from the same XML source are under constant development. Besides the earlier mentioned ways of publishing tables of contents and the references directly from an XML full-text file, we are also exploring possibilities of publishing software to create documents from XML. The recent publication of a print version of the university catalogue from an XML source has been a step forward in the realisation one XML source for all ways of publication.

A prototype of a web based work flow system which evolved out of a project also related to electronic publishing will soon be launched. Initially, the system will be used in the process of creating documents, such as student theses and articles for journals. These processes can be quite complex where copies of documents are sent back and forth via e-mail between authors and reviewers. However, these processes can be facilitated by an application which takes care of the document and information flow. Therefore we are considering to integrate this system into DiVA to facilitate the publishing and ingest work flows.

Summary

DiVA has been operational for over three years now, and the next generation is about to be born. The coming version will be entirely based on open source software and will incorporate new modules for storage, indexing and searching. Furthermore, applications facilitating the publishing work flow and the metadata ingest will be integrated into DiVA. The current version of the XML schema representing the DiVA Document Format will be replaced by a complex and more flexible XML schema. With this new schema it will be much easier to describe different digital and physical objects. It will be possible to extend the object specific metadata with rights metadata, administrative metadata, and file format metadata. The format will be designed to be used as an archival package format for packages containing metadata and digital objects in data streams within the same XML file.

²⁴DIDL is part of the MPEG-21 standard.

²⁵METS specification. <<u>http://www.loc.gov/standards/mets/</u>>

²⁶Pronom home page. <<u>http://www.nationalarchives.gov.uk/pronom/</u>>

²⁷GDFR home page. <<u>http://hul.harvard.edu/gdfr/</u>>

²⁸Fedora home page. <<u>http://www.fedora.info/</u>>

Forming a consortium to ensure future development and the operation of DiVA has shown to be a success. This consortium creates specifications of requirements and further develops DiVA in collaborative way, an approach also used in projects such as Sakai, OSPI and Kuali.²⁹ To share knowledge, experience and costs is of great value for the ongoing and future development of Di-VA.

Appendix

University	Home page	DiVA
University of Gävle	http://www.hig.se	http://www.diva-portal.org/hig/
Jönköping University	http://www.hj.se/	http://www.diva-portal.org/hj/
Karlstad University	http://www.kau.se/	http://www.diva-portal.org/kau/
KTH, Stockholm	http://www.kth.se/	http://www.diva-portal.org/kth/
Mälardalen University	http://www.mdh.se/	http://www.diva-portal.org/mdh/
NTNU, Trondheim	http://www.ntnu.no/	http://www.diva-portal.org/ntnu/
University of Skövde	http://www.his.se/	http://www.diva-portal.org/his/
Stockholm University	http://www.su.se/	http://www.diva-portal.org/su/
Södertörn University College	http://webappo.web.sh.se/	http://www.diva-portal.org/sh/
Umeå University	http://www.umu.se/	http://www.diva-portal.org/umu
Uppsala University	http://www.uu.se/	http://publications.uu.se/
Växjö University	http://www.vxu.se/	http://www.diva-portal.org/vxu/
University of Aarhus	http://www.au.dk/en	http://diva.statsbiblioteket.dk/au/
Örebro University	http://www.oru.se	http://www.diva-portal.org/oru/

Two members of the consortium are missing in the table, because their DiVA installations are planned for the summer of 2006.

²⁹Brooks, Lois (2004)

Figure 1: Metadata and publishing work flow



Figure 2: Resolution Service



Bibliography

- 1. Brooks, Lois. "Values of Community Source Development", Syllabus, September 2004, <u>http://www.-campus-technology.com/article.asp?id=9940</u>
- Klosa, Uwe; Siira, Erik; Andersson, Stefan: "The DiVA Publishing System Uncovered: XML Based Technologies for Advanced Publishing Services". ELPUB 2005, the 9th ICCC International Conference on Electronic Publishing, Leuven, Belgium, 8-10 June, 2005. <u>http://elpub.scix.net/cgibin/works/Show?_id=220elpub2005&sort=DEFAULT&search=year%3a2005%20type%3aposter&hi ts=9 and http://epc.ub.uu.se/files/E-Poster.pdf
 </u>
- 3. Moe, Marianne; Müller, Eva: "ETDs Cooperation within three Nordic countries". ETD 2004: Distributing knowledge worldwide through better scholarly communication, Lexington. <u>http://www.uky.edu/ETD/ETD2004/abstract5.html#moe</u> and <u>http://www.uky.edu/ETD/ETD2004/muller/muller moe final.pdf</u>
- http://www.uky.edu/ETD/ETD2004/muller/muller_moe_final.pdf 4. Müller, Eva: "DiVA - Academic Archive Online". DigiCULT.Info: Issue 9, 2004, 22-25. http://www.digicult.info/downloads/digicult_info_9_xs.pdf
- 5. Müller, Eva: "Access now and in the future: Infrastructure supporting access and preservation of documents in the long term". Joint Workshop on Electronic Publishing jointly organised by Delos, SVEP and ScieCom, Lund, Sweden, 14-15 April, 2005. http://epc.ub.uu.se/files/e_muller.ppt_
- 6. Müller, Eva: "Current state of activities and trends in the ETD's development in Scandinavia". CNI Spring 2005 Task Force Meeting, Washington, D.C., 4-5 April 2005. http://epc.ub.uu.se/files/040505CNI.pdf
- Müller, Eva: "Long-term Access of Theses and Dissertations: The SVEP Project in Sweden". ETD 2004: Distributing knowledge worldwide through better scholarly communication, Lexington. http://www.uky.edu/ETD/ETD2004/abstract5.html#muller and http://www.uky.edu/ETD/ETD2004/muller/muller final.pdf
- http://www.uky.edu/ETD/ETD2004/muller/muller_final.pdf
 8. Müller, Eva: "Persistent identifiers, long-term access and the DiVA preservation strategy". ER-PANET Seminar on Persistent Identifiers, Cork, Ireland, 17-18 June 2004 . http://www.erpanet.org/events/2004/cork/presentations/PID muller.pdf
- http://www.erpanet.org/events/2004/cork/presentations/PID_muller.pdf
 9. Müller, Eva; Andersson, Stefan; Klosa, Uwe; Hansson, Peter: "Metadata Workflow Based on Reuse of Original Data". ETD 2003: Next Steps Electronic Theses and Dissertations Worldwide, Berlin. http://www.hu-berlin.de/ETD2003/userinfoviewpaper.php4?paper_id=15
- Müller, Eva; Klosa, Uwe; Andersson, Stefan; Hansson, Peter: "The DiVA Project Development of an Electronic Publishing System". D-Lib Magazine, (9)2003:11. <u>http://www.dlib.org/dlib/november03/muller/11muller.html</u>
- 11. Müller, Eva; Klosa, Uwe; Andersson, Stefan; Hansson, Peter; Siira, Erik: "Archiving Workflow Between a Local Repository and the National Archive". ECDL 2003, Trondheim. http://epc.ub.uu.se/files/archiving_ECDL_2003.pdf
- 12. Müller, Eva; Klosa, Uwe; Hansson, Peter; Andersson, Stefan; Siira, Erik: "Using XML for Long-term Preservation : Experiences from the DiVA Project". ETD 2003: Next Steps Electronic Theses and Dissertations Worldwide, Berlin. <u>http://www.hu-berlin.de/ETD2003/userinfoviewpaper.php4?paper_id=39</u>
- Müller, Eva; Sandgren, Frida; Andersson, Stefan; Klosa, Uwe; Hansson, Peter: "DiVA publishing system: The community's collaborative development approach". ELPUB 2005, the 9th ICCC International Conference on Electronic Publishing, Leuven, Belgium, 8-10 June, 2005. http://elpub.scix.net/cgi-bin/works/Show?_id=213elpub2005&sort=DEFAULT&search=year%3a2005&hits=69 and
- http://epc.ub.uu.se/files/ELPUBfinal.pdf
 14. Sandgren, Frida; Klosa, Uwe: "Basic content markup of scientific texts as a supportive feature in electronic publishing XML as a format for advanced functionality". Joint Workshop on Electronic Publishing jointly organised by Delos, SVEP and ScieCom, Lund, Sweden, 14-15 April, 2005. http://epc.ub.uu.se/files/sandgren.ppt

Internet sources

- 1. ARK: <u>http://www.cdlib.org/inside/diglib/ark/</u>
- 2. DiVA application profile: <u>http://epc.ub.uu.se/diva-app-profile/</u>
- 3. DocBook specification: <u>http://www.docbook.org/</u>
- 4. DOI: <u>http://www.doi.org/</u>
- 5. Dublin Core: <u>http://dublincore.org/</u>
- 6. FBFR: <u>http://www.ifla.org/VII/s13/frbr/frbr.htm</u>
- 7. Fedora: http://www.fedora.info/
- 8. GDFR: <u>http://hul.harvard.edu/gdfr/</u>

- 9. Handle: http://www.handle.net/
- 10. Marc: <u>http://www.loc.gov/marc/</u>
- 11. METS: http://www.loc.gov/standards/mets/
- 12. MODS: http://www.loc.gov/standards/mods/
- 13. MPEG-21: http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=41112&ICS1=35&IC S2=40&ICS3=&showrevision=y&scopelist=CATALOGUE
- 14. OASIS Open Document Format: http://www.oasis-open.org/committees/office/
- 15. Open Archives Initiative: http://www.openarchives.org/
- 16. Pronom: http://www.nationalarchives.gov.uk/pronom/
- 17. Resolution Service: http://urn.kb.se/start
- 18. RSS 2.0 specification: http://blogs.law.harvard.edu/tech/rss
- 19. URN: http://en.wikipedia.org/wiki/URN
- 20. XML specification: http://www.w3.org/XML/
- 21. XML schema specification: http://www.w3.org/XML/Schema
- 22. XSL specification: http://www.w3.org/Style/XSL/