XML Based Technologies for Advanced Publishing Services: The DiVA Publishing System Under the Hood

Uwe Klosa, Erik Siira, Stefan Andersson

Uppsala University
Electronic Publishing Centre
Dag Hammarskjöldsväg 1
75237 UPPSALA
Sweden
{uwe.klosa | erik.siira | stefan.andersson}@ub.uu.se

Abstract

Within DiVA (Digital Scientific Archive) XML is used for storage and archiving, for inter-application-communication, for creation of web pages, and for the export of metadata. Many parts of the publication work flow (print and electronic), most parts of the archiving workflow, and the metadata workflow are supported by the use of XML. All metadata are stored in XML, more specifically in XML files with UTF-8 encoding. The format used in the XML files is defined by an XML schema that was developed by the Electronic Publishing Centre (EPC) at the Uppsala University Library. This format, the DiVA Document Format (DDF), was created to achieve the storage of the full-text content—not including multimedia like images, films and sounds—and its metadata in a single XML file. Another purpose was to obtain a rich format, which would meet our requirements, such as holding sufficient metadata and the content of a document, and which would be compatible with most standard formats or de-facto standard formats. Furthermore, this format had to be extensible, and usable within an archiving workflow.

This poster will show how XML is produced from the full-text content and with the services that are built and are planned to be built for both document authors and users of the DiVA web pages.

In order to create an XML format from full-text files, the authors have to use templates in the word processor of their choice. It is important that these templates can easily be mapped to the required XML format. This can involve the necessity of a transformation of a proprietary XML format, exported from word processors, to the required format. In the DiVA system this is done from the Open Office XML to DDF.

An XML format supporting both metadata and full-text content has several benefits for the printing workflow, the archiving workflow and the dissemination of publications and their metadata. The printing workflow can be supported by the just in time creation of title or abstract pages. The end user, who wants to get information about a publication, can be provided with all metadata. In addition, the user can be offered important information like table of contents or reference lists directly on the web page, without any need to open the often large full-text files.