Save as XDiML (DissertationMarkupLanguage)

Writing and Converting digital Theses and Dissertations using OpenOffice

Sabine Henneberger
Matthias Schulz
Humboldt University, Computer and Media Service
shenneberger@cms.hu-berlin.de
matthias.schulz.1@cms.hu-berlin.de
Erwin Schrödinger- Zentrum, Rudower Chaussee 26, 12489 Berlin
http://edoc.hu-berlin.de
Keywords: XML, OpenOffice, Conversion

Abstract

For 5 years now, doctoral candidates of Humboldt- University at Berlin can choose the digital publication as one option to publish their dissertation. The Electronic Publishing Group provides stylesheets for Microsoft Word, WordPerfect and Staroffice/Openoffice in order to allow the authors to structure their digital documents semantically. It is neccessary to prepare documents that way, because they will be converted into an XML format afterwards. The XML document then is an instance of the xDiML.DTD (DissertationMarkupLanguage). This DTD has been developed within several electronic publishing projects at Humboldt University.Since 1997 an SGML/XML- based concept for the long term preservation of digital publications has been used, in order to store digital documents in a media neutral archival format, to use the possibilities of a structered retrieval within the semantic structures of documents, and to enable an automated production of different information products (like PDF format for print, HTML for WWW layout, metadata for use within different retrieval networks).In order to increase the efficiency of OpenOffice with it's end format XML, a document style sheet and a filter for Openoffice was developed at the Computer- and Media Service of Humboldt University. Authors are enabled to produce their digital dissertation using this document style sheet and to save those within an XML based format. At the same time the Electronic Publishing Group uses OpenOffice as a conversion tool to convert Microsoft Word documents produced with the digital dissertation stylesheet into XML.

This talk will focus on demonstrating the needs for a dissertation.dtd and it's structure. Secondly the converter "save as xDiml" will be presented and the advantages and disadvantages of producing a complex document like a dissertation using OpenOffice will be discussed.

Expectations to electronic Publishing

If we speak about electronic publishing, we have different expectations about that. We have different views. The author view, with creating and editing. The users view with availabilty and retrieval of the electronic documents. The view of computing center with the problem of long term preservation and view of publishers with quality control. All this views have an effect on the publishing workflow. The main points of the publishing workflow are creating, archiving, retrieval and problems with them.

Creating an electronic document

What do the authors demand? They wish modern text processors with support for their tools for multimedia applications. They want to get a guarantee of integrity and authenticity of their documents. The theses and dissertations should be long term preserved an available for worldwide access. The authors want to create fast their theses and the publication time should be short from submission to publication. But, the libraries need an electronic document, which can be archived and be available for a long time. Therefore we use XML for archiving.

The XDIML

XDiML stands for DissertationMarkupLanguage in XML. XDiML is a DocumentTypeDefinition. A document type definition is a requirement for structuring a dissertation. The first DTD for thesis and disseration was developed from Yuri Rubinsky and Neil Kipp. The DiML was developed out of the ETD-ML of Virginia Tech in 1997. This version was redesigned into XML in 2002.

The DTD has a document structure like book or tei dtd. The root element is etd (electronic thesis and dissertation). The structure of DTD is modulbased like TEI-DTD.

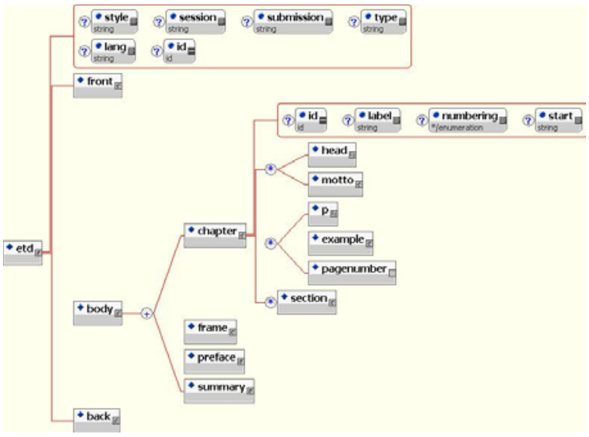


Figure 1: XDIML view with main elements: front, body and back

What we need to get XDIML

I think, that the authors need a modern text processor. OpenOffice is such a modern text processor. OpenOffice is an open source project. But, what we need to get XDiML? The electronic publishing working group of humboldt university developed a template for theses and dissertations with a new menu *Dissertation*.

The authors need templates for formating. All essential templates or styles are collected in the menu *Dissertation* (thesis). The order corresponds to the structure of a thesis. (titlepage = <front), headings... = <body>, appendices = <back>.

The main intension for the template are:

- The students format the specific content. They can also format the semantic terms with specifc styles for intance the titlepagestyles. If the students use the template, we can transform the thesis to XDiML-Format.
- The second aspect is: the template helps the authors to create and format thier thesis and dissertation.

The template disseration is the first step to XDiML format.

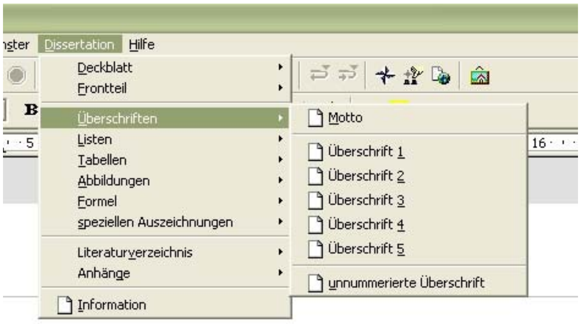


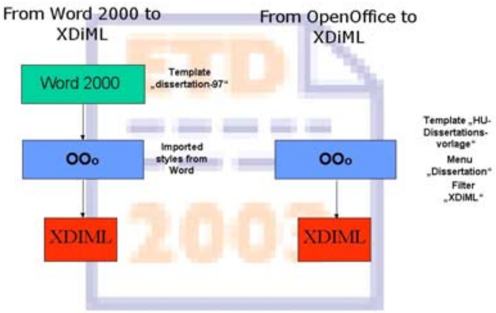
Figure 2: The dissertation *menu in Openoffice with dissertation template*

From Word or OpenOffice to XDiML

On the one hand, we use OpenOffice for converting ETDs created in Word 2000 to xml.

The problem with Word 2000 is, that there is no converter available to xml . Provided, that the document was created with the correct template *dissertationen-97* for ETDs, we open the word document in OpenOffice and all styles will be imported.

On the other hand we create the ETD directly in OpenOffice, using the OpenOffice template *HU-Dissertationsvorlage*. *HU-Dissertationsvorlage* and *dissertationen-97* contain the same styles. In both cases we have now to convert the OpenOffice text document into xml, or more precisely into the XDiML structure.



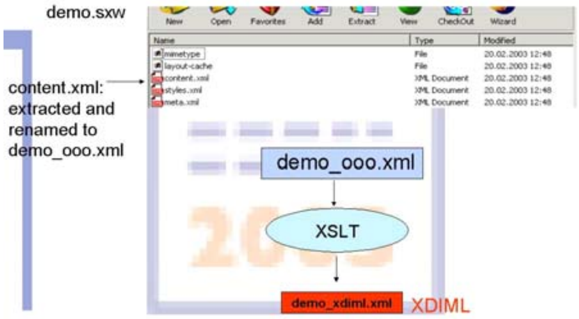
Creating ETDs in OpenOffice

For creating ETDs directly in OpenOffice we use 3 supplements.The first is the already mentioned **template** "HU-Dissertationsvorlage", what is required because here the styles are included. The second supplement is the **menu** "Dissertation", which can be used optionally.

The third is the special output **filter** XDiML, which can also be used optionally.

The OpenOffice Text File Format

Actually the OpenOffice Text file (extension sxw) is an archive. This becomes visible if you open the text file with the program zip, for example. Now one can see the single components of the text file. The component content.xml is the component, we are mainly interested in. You can extract this file and open it with an XML-Editor, in our case we use SPY of Altova. Here you can see the xml structure of the file. For better identification we rename the file to demo_ooo.xml. The goal is to transform this xml file into the XDiML structure (demo_xdiml.xml). XSLT, the Transformation of xml files with Extensible Stylesheet Language, is the appropriate tool for this purpose.



Name	Type	Modified
mimetype	File	20.02.2003 12:48
layout-cache	File	20.02.2003 12:48
content.xml	XML Document	20.02.2003 12:48
styles.xml	XML Document	20.02.2003 12:48
meta.xml	XML Document	20.02.2003 12:48

How to get the XDiML File?

For the transformation from demo_ooo.xml to demo_xdiml.xml an XSLT script is used.But the structures of input and output file are very different. Besides many other details two problems must be solved:

(A) Assigning: The text content (included in element office:body) has to be connected with the appropriate style. There are elements with the correct style already in the attribute text:style-name and there are elements with an attribute text:style-name like "P1". For this element we must search for the appropriate style in another part of the document, inside of element office:automatic-style.

(T) Transformation:The structure of the input is almost flat. Headings of chapters and subchapters are siblings, e.g. on the same level. The structure of the output file is much more nested. To change the first structure into the second structure is a very complex transformation.

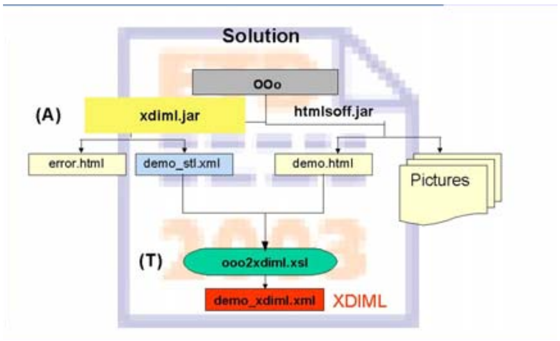
If (A) and (T) are accomplished with one XSLTperformance problems occur for almost every tested dissertation. For this reason we divided the XSLT into two parts, first (ooo2dimlstyles.xsl) realising (A), second (ooo2xdiml.xsl) realising (T).

From the first stylesheet ooo2dimlstyles.xsl we build the Java Archive xdiml.jar and establish it as a filter for OpenOffice. We can run this stylesheet also from the command line using a parser. An error list will be created, which contains all styles not allowed in a dissertation. The pictures of the document are extracted saving the OpenOffice text file as HTML.

Now it is possible, to process larger dissertations. But there are still some problems left.

If we use the first stylesheet as a filter, an error list file is not created, instead of this the error list is included inside the output file demo_stl.xml. The time performance is not satisfying. For larger documents, e.g. more than 1,5 MB we obtain an empty output file. There are no problems in case of starting the first stylesheet from the command line.

If the document is large, e.g. it containes more than 3000 table cells, it is not possible to process it in one single step under Windows Operating System. The processing of a document with 2400 table cells took 17 min with Windows. If we run the stylesheet in contrast under Linux Operating System, any document used so far was processed in one step and 4 to 5 times faster.



Conclusion

For special document types we need filters and templates for OpenOffice. The user needs help: tools (menu), lessons, manuals and support. The conversion is not an one way process. It is a very complex process with different tools or even a process with different parts and file formats.