High quality electronic publishing in universities using XML - the DiDi priciple (Dissertations project at Humboldt-University Berlin)

Susanne Dobratz, Peter Schirmbacher and Matthias Schulz

Humboldt-University Berlin Computer and Media Center

Unter den Linden 6 10099 Berlin; Germany
<dobratz, schirmbacher,matthias.schulz.1@rz.hu-berlin.de>

Introduction

The number of electronic documents, which are to be handled, archived and made accessible to the public by university libraries is highly increasing. Using an SGML/XML-based publishing concept enables university libraries not only to encourage their authors to use new publishing concepts, but also to set up information and document management services with higher quality. There are two main arguments for using an SGML/XML-based publishing strategy for theses and dissertations at Humboldt-University at Berlin: 1. Archiving and 2. Retrieval and the Possibility for Knowledge Management.

Archiving

Archiving electronic texts and documents is a highly discussed topic during the last years. Here two major aspects have to be taken into consideration: first, the question for the hardware, which is used for saving documents and second, the question for an appropriate document format. Answering the question for the document format, which should be used for archiving, the following points seem to be essential to be considered:

- to guarantee a long term preservation for 10 years and more
- the availability of the document format on different hardware plattforms
- the possibility to convert into several document formats without loss of information or data and therefore the possibiliy to choose a presentation format
- standardization of the archive format by an independent international consortium
- the possibility of inclusion of multimedia objects

Retrieval and Knowledge Management

Searching within collections of digital publications is often performed by a fulltext search engine in combination with a search within bibliographic metadata, such as title, author, keywords according to an classification schema and abstracts. Using semantic and semisemantic parts of digital documents offers new perspectives for a more detailed retrieval, which leads to search results, containing higher value of information. Arguments for SGML/XML are e.g. the possibilities :

- of using document structure and semantic tags for retrieval
- for detailed search facilities
- for automated cataloging
- for information extraction (e.g. citation index) and the use knowledge management functionalities
- of deriving highly structured information, which is more valuable than information provided in standard text documents, e.g. in PDF

The call back of a search for a specific term in a fulltext database includes a lot of hits, which are not really relevant for the user. Searching, e.g. in headings of captions for this word decreases the number of hits and leds to information, which is more important to the user, because it is scientific practice to use scientific terms or keywords in caption headings, in order to explain them in the following section.

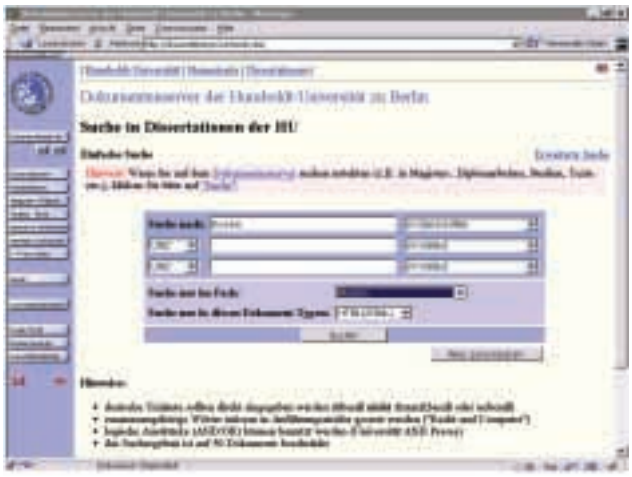

Figure 1: Retrieval Interface of Humboldt-Universities Document Server

The "Digital Dissertations" project at Humboldt-University Berlin was launched in September 1997 by the university library and the computing center of the Humboldt-University in Berlin, with the main aim to built up a secure and robust digital library for university publications using an SGML/XML-based publishing concept. Because acess to those university publications as theses and dissertations is free, it is a good field to investigate new concepts as nearly no copyright laws have to be considered. The objectives are on questions of human-computer-interaction, authors support, markup (SGML/XML), retrieval and workflow issues.

The decision for using an SGML/XML based publishing strategy was done by using experiences made in the German Medoc-Project. as well as own evaluations.

Semantic and semisemantic tags, found in content and structured analysis of several theses and dissertations, are used for document description. Comparing the developed DiML-DTD to others as e.g. used at the technical University of Helsinki. University of Lyon2, Montreal, University of Iowa and others, leds to the conclusion, that at the actual state, people seem to habe the same intensions, when thinking about those kind of documents. The differences have to be handled using new XML-Technologies as namespaces and schemes. After the investigation of typical document structures led to a DiML-

DTD (Dissertation Markup-Document Type Definition), which was adopted from a ETD-DTD by Viginia Tech, seewww.ndltd.org, a workflow concept was designed and the systems is now used in practice.

Trying to establishing an conversion workflow model brought up the decision to built up a high quality support for authors into our concept, because investigations of author's capabilities to produce structured documents had come out with the results, that the knowledge concerning the correct usage of simple wordprocessing systems such as Microsoft Word or WordPerfect, even of the LaTeX-System, is barely developed. Important for the project is, that the developed workflow model could partly be established within the university library and done by librarians or staff of the computing center. Therefore a robust solution for archving digital publications has been developed. This includes the usage of digital signatures according to the German Signature Law as well as the investigation of document conversion techniques and workflow models, not only for text elements.

Digital Libraries which using their document servers as long term electronic archives will not make printed information dispensable. On the contrary for users of these information systems the desire for printed documents is increasing. In most cases this desire often focuses not on the whole document as such, but on particular parts of it like chapters, citations and so on. For that reason our printing on demand project aims on the development of a technology which allows the users to print the wanted part of a certain document only.

While the amount of electronic available literature has been increasing during the last years due to the set up of digital libraries, the „technical" quality of the documents has not improved meanwhile. PDF has been widely used for archiving digital documents, although it lacks of standardization and possibilities for structure encoding. It can only be viewed using spezialized browsers (Adobe Acrobat Reader). Content analysis by markup languages like SGML or XML and a semantic encoding of pieces of information are excluded in projects which are only dealing with PDF and therefore propagate a classical print oriented way of thinking. But SGML/XML-based techniques have been approved for information processing and extraction in several fields.

A lack of the right skills by publishers may be the reason for this defizit. Often authors have problems and difficulties using adequate/ appropriate techniques for a computer aided content analysis. One major task of this project will be the concentrated support for authors with sylesheets, courses and online materials.

The Scope of a new project by the university library and the computing centre of Humboldt-University Berlin in close cooperation with the State and University Library of Lower Saxony, Göttingen is to connect two document servers (serving as examples for others) with one retrieval interface, which supports metadata search (Dublin Core) and fulltext search

- the installation of a complete workflow (starting from author support via preparation and archiving of digital documents and image digitazations up to the distribution of those objects and a printing o demand service on both places: Berlin and Göttingen)
- the measurement of acces data and network usage
- the increase of the „technical" quality of documents using an XML-based approch

The server can be visited at:

- http://dochost.rz.hu-berlin.de/
- Project descriptions are at: http://dochost.rz.hu-berlin.de/epdiss
- or http://dochost.rz.hu-berlin.de/proprint