Electronic Publishing '01 –
2001 in the Digital Publishing Odyssey
A. Hübler et al. (Eds.)
IOS Press, 2001

High quality electronic publishing in universities using XML - the DiDi principle

Peter Schirmbacher, Susanne Dobratz, Matthias Schulz

Humboldt-University Berlin, Computing Centre
Unter den Linden 6
10099, Berlin, Germany
http://dochost.rz.hu-berlin.de
(dobratz | schirmbacher | matthias.schulz.1 @rz.hu-berlin.de)

The number of electronic documents, which are to be handled, archived and made accessible to the public by university libraries in cooperation with other institutions, is highly increasing. Using an SGML/XML-based publishing concept enables universities not only to unlock their information and research resources but also to set up new services, like a document management or a printing on demand with higher quality as by using conventional publishing concepts. There are three main arguments for using an SGML/XML-based publishing strategy not only for theses and dissertations at Humboldt-University Berlin but also for university publications like conference proceedings, public readings, technical reports and other material, that is subsumed under the term "Grey Literature": 1. Archiving, 2. Retrieval and 3. Reusability of structured documents. This paper will describe the policy and technology that stands behind the local electronic publishing concept.

1 Enabling technologies

"SGML and XML are markup languages for presenting information as structured documents. They are often called "enabling technologies". To cite Oxford English Dictionary: *Enable, v. To Supply with the requisite means or opportunities to an end or for an object.* Markup Languages do not, by themselves give off-the-shelf answer to questions related to electronic publishing... They provide an opportunity. An opportunity to take a look at the information from a more abstract level. Not as pages or lines, but as packets with a label. When information is "tagged", it means you have classified its distinct logical elements. It is like turning your information into a miniature library, searchable, retrievable, useable.[1]

When talking about an SGML/XML-based publishing concept, there is always the question for the costs of material, tools and workload and its relation to the result. One can barely see that the technology behind a book or a document server is based upon those markup languages. And as long as we cannot provide new and better services, that motivate users to use those digital libraries, those questions have to been answer. Thinking in terms of "Digital Libraries" those new service demands include an availability of services 24 hours a day at 7 days a week independently from the location of the user, giving free access from the desktop to an unlimited repository and a sustainable digital

[1] Tuija Sonkkila in her invitation to the RAJU2001 Seminar, 23.4.2001 Otaniemi, Espoo, Finland. (http://kirjasto.oulo.fi/raju2001/)

archive. More than ever do the typical responsibilities of a library become important: the acquisition, exploitation and the offer of access to information resources. Building systems that provide a better user-oriented retrieval and a higher quality of access to digital literature at low costs becomes therefore essential, if libraries want to survive in an information society.

2 Arguments for an SGML/XML-based publishing concept

2.1 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 questions 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 platforms
- the possibility to convert into several document formats without loss of information or data and therefore the possibility to choose a presentation format
- standardization of the archive format by an independent international consortium
- the possibility of inclusion of multimedia objects in their native format

2.2 Retrieval and Knowledge Management

Searching within collections of digital publications is often performed by a full text search engine in combination with a search within bibliographic metadata, such as title, author, and keywords according to a classification schema and abstracts. Using semantic and semi-semantic 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 and new services like a printing on demand
- for detailed search facilities
- for automated cataloging
- for information extraction (e.g. citation index) and the use knowledge management functionality
- 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 full text 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 led 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.

The local 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[2]. The

[2] see Ohst, Daniel: Dateiformate für das elektronische Publizieren, Studienarbeit am Institut für Informatik, Humboldt-Universität zu Berlin, Berlin, 1998, http://dochost.rz.hu-

experiences and tools were also taken for evaluation and distribution within the German Dissertation Online project, which was funded by the Deutsche Forschungsgemeinschaft from April 1998 until October 2000.[3]

3 Electronic publishing

There are different views on electronic publishing:

View of Authors: Authors want to create and edit electronic publications in their native text-processing environments. For them publishing means disseminating their research results. So they first of all focus on the content rather than on the writing behaviour itself. To publish their documents very fast and worldwide is often the main argument to choose an Internet publication. But often questions as how to ensure intellectual property rights or copyrights often withhold them from choosing this method of publication.

View of Users: Users want to access digital publication, regardless of time and location. The possibilities of retrieving only the wanted rather than all available information for a certain topic seems to be rather essential for them. Another interest of this group is the possibility to prove the identity, integrity and authenticity of documents on servers. Did the author indicated on the front page really write this paper? Can the publishing date secured in a way, that ensures the originality of the information, that can be linked to inventions or patents that have to be secured? Is it possible to prove whether the author or anybody else has altered this document after publication?

View of Libraries: The main duty of libraries is the acquisition, exploitation, and long term archiving of printed materials. These duties now apply to digital documents as well. Libraries are also in charge of ensuring authenticity and validity of electronic publications.

View of Computing Centres: Computing centres usually approach this problems from a perspective where terms like physical availability of servers, bandwidth of computer-networks, quality of retrieval tools, long term archiving hardware, search machines, and storage capacity play a major role.

View of Publishers Publisher consider electronic publishing first of all as an technological process, that has to be handled in co-operation with several partners, such as authors, lectors and service providers.

Launching an electronic publishing project at a university can therefore only be a joint effort of the different institutions that are involved in a publication process:

Research Institutions as the producers of publications

Libraries as the publishing house, being responsible for dissemination

Computing centres as the institutions that holds most of the knowledge of the technology

Media centres as the institutions that provide knowledge for non-textual enhancements.

Such a concept was taken when launching the "Digitale Dissertations" project at Humboldt-University.

---

berlin.de/docserv/buecher/ohst-daniel/HTML/ or Schirmbacher, Peter: Dateiformate ein zentraler Punkt des elektronischen Publizierens, Vortrag auf dem "Expertenworkshop": Neue Organisationsformen elektronischer Veröffentlichungen: Angebote wissenschaftlicher Bibliotheken, Dortmund, 23.-24.11.1998, http://eldorado.uni-dortmund.de:8080/bib/98/workshop/schirmbacher

[3] see results at http://www.dissonline.de

4 The Digital Dissertations project at Humboldt-University

The project "Digitale Dissertationen" at Humboldt-University in Berlin was sponsored for 3 years (Sept. 1997 - Dec. 2000) by a German government programme for multimedia applications for universities (Hochschulsonderprogramm III). Organised by the University Computing Centre and the University Library, it aimed to build up a digital library of theses and dissertations, access to which is available via the Internet. The project has a number of objectives upon which effort was concentrated. These include:

Addressing security and archiving issues - by the use of digital signatures, electronic time stamping services, and the appropriate use of SGML/XML standards.

Implementing a high quality retrieval system

Ensuring that support is available for authors who are producing their work in electronic format - by providing courses and guidelines as well as special services for scanning and preparing audio, graphical and video information

developing a workflow model for the submission and handling of digital publications within the university - giving considerations to cataloguing, storage and organisation of theses, and the work involved for the staff of the Library and the Computing Centre

This project was extended with the participation within the German "Dissertationen Online" project undertaken by the Initiative of Information and Communication of the German Learned Societies. This project was funded by the German Research Foundation (DFG: Deutsche Forschungsgemeinschaft) for 2 years. The final conference took place at Berlin in October 2000. It worked at a more integrative level and aimed towards a German wide initiative to bring scholarly publications online which are usually lost in libraries, such as dissertations, diploma and master theses. Through the joint work of 6 academic disciplines (mathematics, physics, chemistry, educational sciences, computer science and libraries as the State and University Library Lower Saxony (SUB Göttingen) and the German National Library (DDB: Die Deutsche Bibliothek)) which took place at several locations (Duisburg, Oldenburg, Erlangen, Berlin, Göttingen, Frankfurt) was highly successful in Germany and elsewhere. So a tight co-operation with the Networked Digital Library of Theses and Dissertations (NDLTD[4]), set up by Edward Fox from the Virginia Polytechnic Institute and State University, USA, was established.

5 Developing a electronic publishing concept for Humboldt-University Berlin

By now the project has moved out of the project status into a "normal" workflow between those two institutions. The developments are likely to guide the university to establish an electronic publishing house, but as those decisions have not been made yet, the organisational and technical concept is considered as the most important part. By May 2001, Humboldt-University has tested and established three major publishing lines, that are cover different fields of "grey literature":

Electronic dissertations and professorial dissertations,

Proceedings of the EUNIS2001 conference

Public readings of the university.

For those three different applications different publishing strategies were developed and used.

[4] http://www.ndltd.org

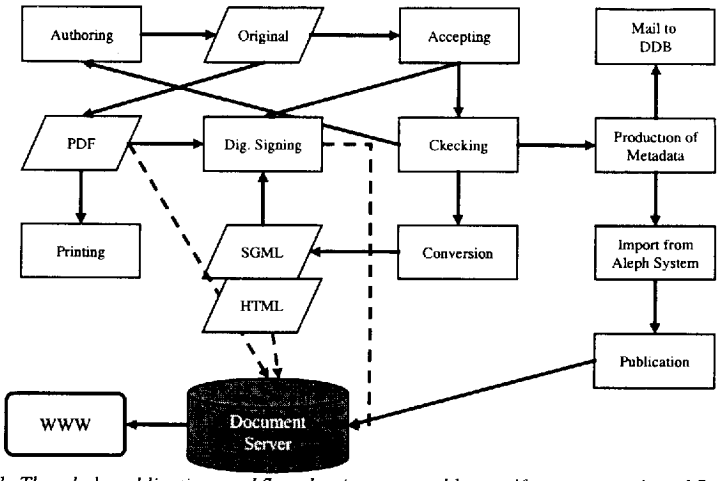
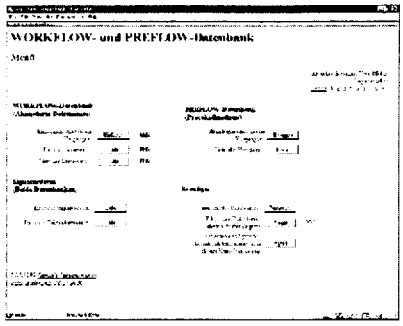
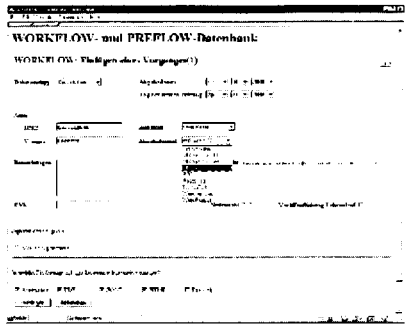
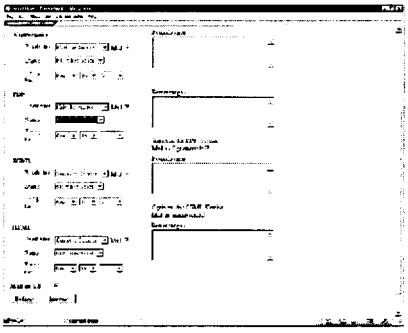


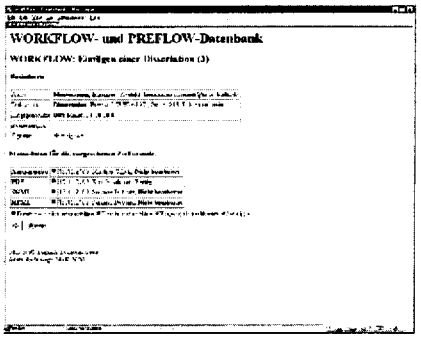
Figure 1: The whole publication workflow that is supported by a self-programmed workflow system

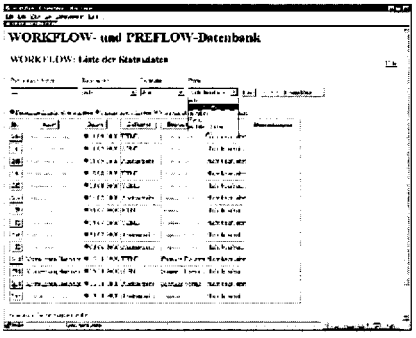
A very sophisticated task was to establish a common workgroup of staff from both participating institutions: library and computing centre and to develop a workflow. The workflow is supported by a workflow system that is easy to use for all parties and allows a detailed control of subtasks and contains a messaging and alerting system for the participating persons. Throughout long discussions and by reorganising of parts of old workflow within both institutions it took nearly 3 years to reach a result, that works independently from the actual project group. The underlying systems is based on a Sybase database, and maintained in the computing centre. The major workflow components are WWW-interfaces programmed using PHP4. So platform independence and an IP-based security strategy was reached by using the HTTP-protocol as basis.

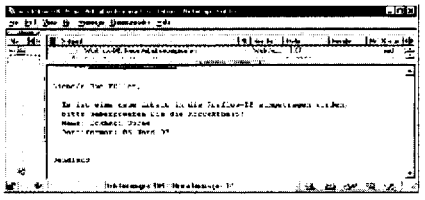












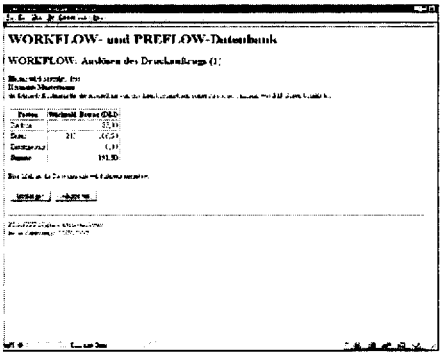


Figure 2: Screenshots of the workflow database user interface

5.1 Electronic dissertations

Why dissertations? In Germany dissertations have to be published by law. A doctoral candidate has to submit his dissertation using one or the other method to provide his or her dissertations to the university library in order to gain the doctor's degree. Usually the doctoral candidate fulfilled this requirement by providing 40 up to 120 paper copies (depending on the department) to the library or to provide a certain amount of microfilm copies to the library or to publish the whole dissertation in an publishing house or as an special issue in an well known journal in his field. All those conventional methods usually put a heavy financial burden upon mostly young scientists. So in 1997 the "Kultusministerkonferenz" (Conference of the Federal Ministries of Culture) decided to advise universities to allow a digital publication as additional method to fulfil the publication duty. Like every university in Germany, this advice had to be put into local law by the university itself. Humboldt-University did so by establishing the legal basis in March 1998, with an addition to all "Promotionsordnungen".

5.1.1 Document Conversion Workflow

The aim to collect a critical mass of documents and to develop a conversion strategy to SGML/XML was soon reached. By April 1998 the basic conversion concept was in place and documents could be converted from Microsoft Word into SGML instances. This technology has been improved throughout the entire project time, so that by now, a nearly automated conversion could be performed, which takes about half an hour for an average dissertation of about 200 pages, including several tables and figures.

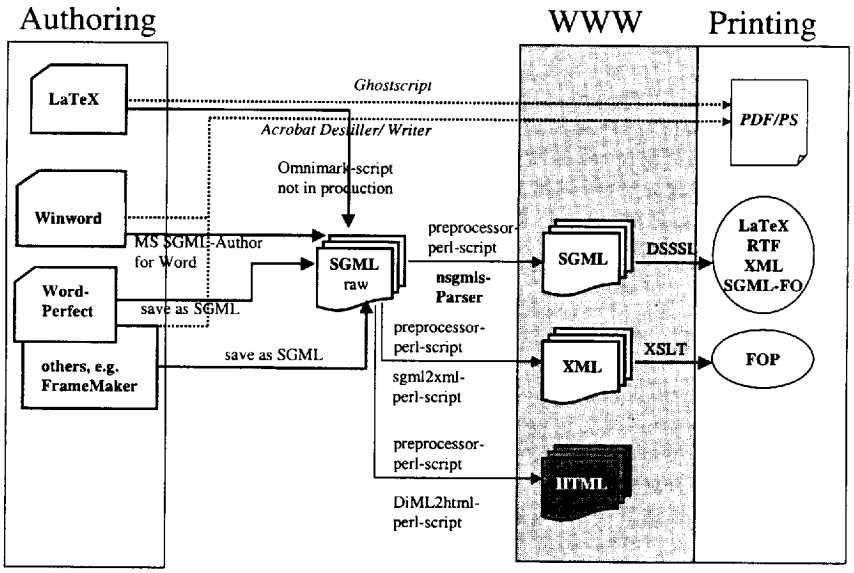


Figure 3: Schematic view onto the conversion process
(FOP: Formatting Object to PDF engine;
SGML-FOT: SGML: Formatting Object Tree;
RTF: Rich Text Format)

The usage of Markup Languages like SGML/XML is not an easy task. First, a DTD (document type definition) giving the structured demands for the documents was designed. Instead of starting from the spread, the project used developments from the Virginia

Polytechnic Institute and State University[5] (Virgia Tech). This ETD.dtd was adopted to meet the German demands and renamed into DiML.dtd (Dissertation Markup Language). The technology of Virginia Tech using a tool "Microsoft Author for Word" was improved for certain details as table model, bibliography model, etc. and put into practice.

The first idea within the project was to provide an SGML/ XML -based editor to the authors in order to get "tagged" documents. A survey that was conducted in early 1998 appeared with a result that was not surprising. Potential authors wouldn't accept a higher effort, e.g. to learn the usage of a new text formatting system or XML-based writing systems in order to publish electronically. So supporting authors within their native word processing system in order to obtain pre-forms of achievable originals became a major task for the project. Especially as 75% of the authors used Microsoft Word, the focus turned towards this user group. A high quality authors support, resulting in the offer of a choice of 2 4-hour Word courses, which teach the usage of style sheets and Word features was established. Whereas project staff conducted the first courses they became a stable offer of the computing centre and are now held by staff of the computing centre.

Table 1: Usage of word processing systems at Humboldt-University

Word 6-9	166	75%
LaTeX	47	21%
Corel Wordperfect 7-8	5	2,3%
FrameMaker	3	1,4%
QuarkXpress, PageMaker	0	
Total of submitted dissertations (Jan. 1998 - July 2000)	221	100%

5.1.2 Metadata workflow

The German "Dissertationen Online" (DissOnline.de) project developed a Dublin Core metadata set for electronic theses and dissertations[6] that has been agreed by all German libraries. With this basic technology, a metadata database, covering several functions has been built upon a Sybase system using Java as programming language. It serves the following tasks:

Create catalogue entries for the digital documents and establish an interface to the Libraries Online Public Access Catalogue

Providing access to the digital documents through a highly configurable and reusable WWW-interface, that supports navigation within the archive and search using the Dublin Core fields.

Reporting new documents nearly automatically to the German National Library (Die Deutsche Bibliothek (DDB)), that has the duty to collect all German language literature.

The metadata workflow is an essential part of the whole publication process and is a major part of the workflow supported by the workflow database. Only if the librarian has finished cataloguing the publication, she or he can give the entry free for public access. This online publication than becomes visible for WWW-users and is part of the list of available publications shown in the screenshot below.

[5] http://www.etd.vt.edu

[6] see actual information at http://deposit.ddb.de/metadiss.htm

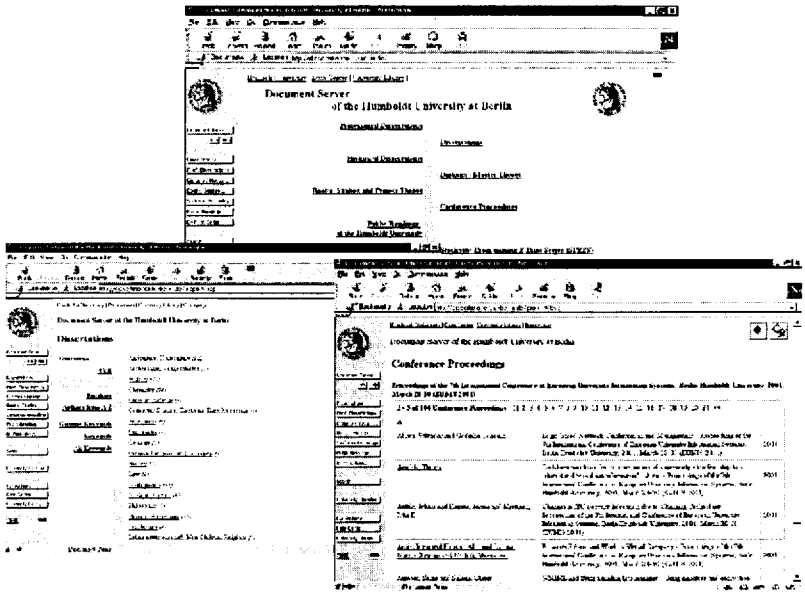


Figure 4: List of available publications, e.g. Dissertations that is automatically produced by the metadata database at the university library

5.2 EUNIS 2001 proceedings

According to the developed workflow, the concept was proved while applying it to another type of documents. The difficulties that arose with the EUNIS 2001 articles for the conference proceedings were the following ones:

The Word style sheet was designed using German version of Word. This was not acceptable at Word systems in other languages, e.g. Russian, Bulgarian, etc. Problems with character sets are in focus as well as the different naming of standard paragraph and character styles.

The Dissertation word template was much too comprehensive for the document type article.

The Word Macros written in WordBasic partially failed to work for several reasons: character sets, naming of styles (the macros were build based on German style names). This was esp. difficult to handle for the bibliography database, that the authors were obliged to use.

The authors had severe problems in using the Word template. Mostly the reason for that was that in contrary to the authors of dissertations who usually undergo a detailed introduction to that topic, those authors were not at all used in structured publishing. So most of the received articles had to be structured by the local publishing group.

Many problems were caused by the inclusion of graphics within the documents. As detailed instructions were given on how to save images, photographs and graphics, most of the authors failed to do so. Painting diagrams in Word was one of the major mistakes that were made.

The conversion technology applied to the conference proceedings was basically the one used for the dissertations (Microsoft SGML Author for Word 97). For the conference proceedings, the digital preliminary printing stage was set in an SGML based styling

system 3B2[7]. So the whole proceedings were printed directly from the SGML source code. In order to accomplish a cross media publishing using the same document sources, the SGML documents were proceeded using a perl-script, so that they became XML documents. Those XML documents were not, as it is the technology for the electronic dissertations, by another perl-script converted to static HTML documents, that could be viewed on the web. They are on-the -fly generated using the Apache[8]/Cocoon[9] XSLT-Server technology.

5.3 Public Readings of Humboldt-University

As Microsoft is not developing their conversion tool further on[10], although it guarantees a high quality of conversion, another additional conversion technology had to be found. Therefor a comprehensive test of FrameMaker + XML 6.0 as conversion tool was conducted. It was tested with public lessons of Humboldt-University, which are basically the entrance readings of professors or readings at special occasions. This document type distinguishes itself from the other through the fact, that there is just one author, who is responsible for the publications. That person works in the research department of the university and usually receives the documents without any formatting as Word files. The previous publishing strategy used QuarXPress with bare formatting and structuring in order to produce a layout for a printed brochure. In order to reuse the document for a high quality internet publication, the decision was made to use FrameMaker as new layout tool because it supports this task in a very professional manner and additionally it is able to deliver also structured versions of the documents. So a complete XML environment has been set up, consisting of:

a new Document Type Definition: OVL.dtd

an FrameMaker+XML SGML application

an EDD (Element Definition Document) for applying structure to a FrameMaker document

a conversion table to map those FrameMaker elements to XML elements according to the new OVL.dtd

a layout style that can be applied to the FrameMaker EDD elements (and for production of the printed / PDF version)

an XSLT an a CSS style sheet for applying a layout to the exported XML document for usage on the WWW (a server side XSM/XSLT processing using Cocoon has been installed)

As a usage of this originally intended conversion technology for the electronic dissertations, this comprehensive test resulted in the realisation, that for a heterogeneous author environment this technology is much too complicated to be used. As the usage of FrameMaker for writing demands much more skills in order to produce pre-structured documents, this technology cannot be recommended for the electronic dissertations.

6 Printing-on-Demand Service with Apache/Cocoon

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

---

[7] http://www.3b2.com

[8] http://www.apacheorg

[9] http://xml.apache.org/cocoon/index.html

[10] The latest version of Word it is available for is Word97. As Word 2000 has the same internal document format (doc) those documents can be converted using Word97.

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.

For the printing on demand component with XML the usage of Apache/Cocoon was chosen. This software uses an XSLT-engine to produce an HTML or PDF-Version on the fly. "The Cocoon Project is an Open Source volunteer project under the auspices of the Apache Software Foundation (ASF), and, in harmony with the Apache web server itself, it is released under a very open license. Even if the most common use of Cocoon is the automatic creation of HTML through the processing of statically or dynamically generated XML files, Cocoon is also able to perform more sophisticated formatting, such as XSL:FO rendering to PDF files, client-dependent transformations such as WML formatting for WAP-enabled devices, or direct XML serving to XML and XSL aware clients. "[11]

As Cocoon does not consist of a printing on demand component (especially a selection feature) a small workaround using different XSLT-style sheets had to be done. The users view, containing an HTML-view onto the actual document includes check boxes, which the user can use to select parts of a specific document. This view is produced by the XSLT-Broker style sheet which calls a default style sheet, that produces HTML (XSLT-Style sheet with option document.xml?format=html) if the user selects certain parts of the document by clicking in the checkboxes. Then by clicking on the "OK" button a perl script (PHP-Choice) is called. This script selects the desired chapters and sections of the document by using XPath-expressions (http://dochost.rz.hu-berlin.de/proprint/bsp/slides.xml?CHAPTER=3 and http://dochost.rz.hu-berlin.de/proprint/bsp/slides.xml?CHAPTER=4) and cuts those parts out of the document and holds them in the main memory. This procedure is carried out by the XSLT-Broker-style sheet that is now been called with the XML-option (document.xml?format=xml). These parts are added to one single XML-document (all in the main memory!) and processed by the XSLT Broker-style sheet either with the print option or the HTML option (document.xml?format=pdf or document.xml?format=html)

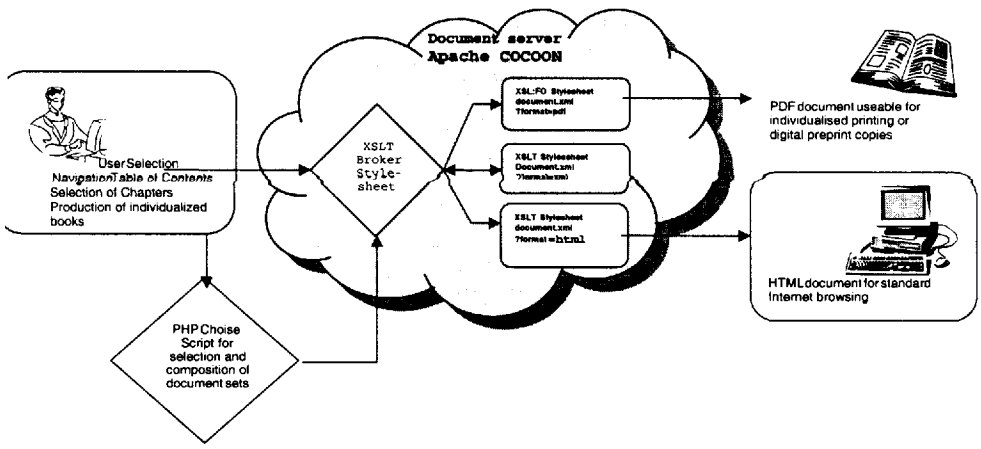


Figure 5: Usage of Apache/Cocoon for Printing on Demand

[11] http://xml.apache.org/cocoon/index.html

7 Further Developments

Looking at the local developments made within the past 3 years, it is now inevitable to establish a broader publishing policy that covers ideally all university publications. Such a policy should also include concepts regarding:

a reviewing system

a payment policy and component

a broad distribution and marketing policy

The education of students in the field of scientific (electronic) publishing from the beginning of their study.

All those ideas can only become reality of not only the service institutions, but also the university management are working together on that issue.

Further developments regarding the technical question of conversion will lead to more support for authors and therefor the investigation into alternative word processing systems, the customisation of XML editing solutions and developments made for a server side automated usage of conversion tools will proceed.