THE PROJECT "THESES ONLINE" PUBLICATIONS BETWEEN SCIENCE AND LIBRARY

Dipl. Inf. Susanne Dobratz
Projekt »Digitale Dissertationen«
Rechenzentrum
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin

Tel.: +49 30 2093 2475
Fax: +49 30 2093 2959
mailto:susanne.dobratz@rz.hu-berlin.de
http://www2.rz.hu-berlin.de/~h0077dfz

Dr. Hans-Ulrich Kamke
DFG-Projekt »Dissertationen Online«
Abt. Pädagogik und Informatik
Philosophische Fakultät IV
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
Tel: +49 30 2093 4177
Fax: +49 30 2093 4198
mailto:kamke@educat.hu-berlin.de
http://www.educat.hu-berlin.de/diss_online

The project "Theses Online" (Dissertationen Online), sponsered by the German Research Foundation (DFG) and initiated by a subgroup within the Initiative of the German Learned Societies for the Advancement of Digital Information and Collaboration ("IuK-Initiative"), started in spring 1998 and was terminated in March with a conference held in Jena, Germany. Funds for a second project year were granted by the DFG with a heavy emphasis on the collaboration with libraries and university computing centres, with research and development running from March 1999 to March 2000.

Among the learned societies involved in the project are chemistry, computer science, education, mathematics, and physics, involving five German universities. Participants in the second proposal are also computing centers, libraries and the German National Library (DDB).

Traditionally, in Germany every graduate student is obliged to publish his doctoral thesis, putting a heavy financial burden on young professionals. Unless the thesis is published by a well known publishing house, theses often are not easily accessible. Furthermore, retrieval by means of bibliographic sources will be combersome, if not impossible. With the advent of digital production, a convincing alternative model is being developed, using the Internet as means of dissipation as well as retrieval, thus making scientific research more productive.

The project evoked intensive communications between learned societies and libraries concerning a special type of scientific publication: theses. The discussions of the last year, which have gone far beyond valuable, but isolated single projects in the past, made meaning and consequences of electronic documents lastingly clear: Archiving and supply of research results laid down in theses do not any longer represent a mere act of administration of the libraries. Rather, under the conditions of modern electronic publication possibilities, archiving and protection of scientific work in electronic form as well as retrieving scientific information via "meta data" from digital sources necessitates the active participation and collaboration between learned societies, libraries and graduate students is indispensable.

The learned societies can bring in their demands regarding the graduation procedures and the search aspects necessary for the individual scientific subject and offer a fast and economic publication form to graduate students, enabling a quick world-wide dissimination of research findings. For libraries, a precise arrangement is necessary, defining the format of documents and meta data for different objectives: retrieval, reading, printing and archiving. The inclusion of the German National Library (DDB) in the project is also necessary, since this library is obligated legally to collect theses of the Federal Republic (also in electronic form) and have them accessible in years hence. Also, cooperation with publishing houses seems necessary. This was discussed at the Jena conference in March 1999.

At a time of rapid development within the area of electronic publication, coordination between the parties involved - faculties, computing centers, libraries, publishers – is indispensable. Coordinating the efforts of different learned societies by agreeing on common interdisciplinary, basic assumptions and by developing mutually acceptable concepts and solutions, will produce synergies and guarantee widespread acceptance.

For the scientific use of theses not only in the humanities and social sciencies, it is necessary to offer not only bibliographical metadata and full text but also structural information for retrieval purposes, such as

1. tables of content
2. headings of tables and graphs
3. reference to important contentwide terms (special index, name index etc.)
4. references (links) to external sources (printed as well as Web sources)
5. bibliography
6. references within the work
7. definitions
8. mathematical / chemical formulas
9. theses / hypotheses

These structural meta data are an integral part of the document and have to be defined by the author himself. At present, this predominantly takes place over formatting the text (e.g. headings, footnotes etc.). In order to be able to use these structural data also for a retrieval, they must be tagged as such by the author, either by the use of a structured language like LaTeX, or by "style sheets" as with WinWord.

The project is structured in several parts that together form an integral approach:

1. Metadata (Prof. Törner, Duisburg University)
   implementation of RDF
   adaptation of existing tools for the use of RDF
   implementation of DTD's

2. Retrieval (Prof. Hilf, Oldenburg University)
   installation of Harvest (broker&gatherer)
   distributed search engines
   searching in mathematical formulas

3. Formats (Dr. Schirmbacher, Humboldt University)
   converters from LaTeX to SGML/XML
   converting of an existing SGML-DTD to XML

4. Multimedia (Prof. Gasteiger, Erlangen University)
   creating of an easy-to-use toolset for libraries
   multimedia theses in medicine, mathematics, physics

5. Support (Prof. Diepold, Humboldt University)
   tutorial system for graduate students
   supporting information for several groups (faculties, libraries, universities etc.)
   web pages and CD-ROM as basic information

6. Libraries (Dr. Niggemann, DDB; Prof. Mittler, Göttingen University)
   test of products from the research groups in libraries
   discussion with libraries
   connect to libraries
   problems of long term archiving
   metadata and RDF

On the basis of the experiences, which were already made within the two projects

"digital theses" (http://dochost.rz.hu-berlin.de/epdiss) and
"theses online" (http://www.educat.hu-berlin.de/diss_online/index.html)

at the Humboldt university Berlin, the following areas are to be discussed:

1. file formats for electronic publishing,
2. authors and science,
3. retrieval possibilities in electronic documents,
4. problems of authenticity and integrity of documents,

File formats for electronic publishing

The question about file formats, which are usable for university libraries for electronic publications and from it the following recommendations became within the last two to three years in most diverse places extensively discussed. Therefore it is referred here to well-known publications in Germany ([Ohst 1999a] [Ohst 1998b], [Schirmbacher 1998b], [DiML-Dokumentation1.0]), and the recommendations of the DDB regarding this topic.[1]

At the Humboldt university[2] chiefly two arguments led to the use of SGML (see [Rieger 1995], [Goldfarb 1990]) and/or XML (see [Behme/Mintert 1998]), to select SGML/XML as the best suitable file format for electronic publications. That is on the one hand the argument of long-term archiving and to the second that of retrieval.

Long-term archiving

By its availability on different hardware platforms, the independence from operating systems as well as its convertibility into other data formats (presentations, print and retrieval formats) without lost of data and the associated freely selectable presentation according to contents as well as the standardisation by an independent, international committee SGML/XML is regarded as the format, which guarantees the best legibility in future decades and is best suited for archiving.

Retrieval

By its defined structure SGML/XML is particularly well suitable for the search in a large set of documents of same type. Thus goal-more exact information searches become possible,[3] since the knowledge structures can be standardized here over the quantity of the documents which can be administered. A condition here is the structuring of the text into semantic and semisemantic units.

Another argument for the use of a SGML based publication concept is the possibility of the use and/or the integration of multimedia elements.

Authors and science

Theses are originally not intended for an electronic publication by graduate students; even today the production of a paper copy is the center of attention; therefore the authors use the same text processing systems for the production of the digital publication as for the production of the printout. The most usual systems are different versions of Microsoft Office, Corel WordPerfect, Star Office or, in special fields, LaTeX. These programs store texts usually in proprietary file formats, which are not in WWW readable formats. In order to provide from these proprietary file formats electronic publications for the use in the WWW, it requires a process, if the electronic

[1] http://deposit.ddb.de/formate.htm, Stand: 03.05.1999

[2] And other universities like the Virginia Polytechnic Institute and State University (http://etd.vt.edu), University of Montreal (http://www.pum.umontreal.ca/theses/), University of Iowa (http://vedavid.org/diss/), Helsinki University of Technology (http://www.hut.fi/Yksikot/Kirjasto/HUTpubl/), University of Lyon 2 (http://iep.univ-lyon2.fr/IEP/Recherche/theses.html)

[3] Similar attempts are made in the Global-Info-Program "Carmen - ein integriertes Hypertext- und Informations-Retrieval-System für digitale Bibliotheken" by Norbert Fuhr, University Dortmund.

publication isn't limited to the production of a PDF or a Postscript file. This process must take place on one hand by the author himself distinguishing certain parts of the text as heading, register term etc. and also assigning some meta data, and on the other hand by the library or the computing centre, if the text is converted into a archive/presentation format. The libraries also catalogue the text according to the appropriate sets of rules, report to the DDB etc.. Parts of these steps can be automated if in the processing of texts the necessary precautions was met: e.g. production of the so called "Katalogisat" from the meta data and the production of a presentation format for the Internet.

If the high claims of quality for retrieval and archiving are keped it is necessary however to train the authors as early as possible and make them familiar with basic questions and problems of electronic publishing.

The work, once invested into the development of certain tools and into the processing of texts itself, is however not end in itself of libraries, computing centres or even the authors. The profound development, which stands among other things at the end of this work, serves also the science by a quicker access to thesis, and better searchability. By the creation of meta informations (registers, tables, formulas) in the research various possibilities of the search, which are in a printed text today only heavily or not at all possible. In addition there is the possibility to integrate not only multimedia elements (sound, picture, video etc..) into a scientific work, but also to search such elements, as soon as these elements are tagged as multimedia.

Retrieval in structured documents

At present a usual search practice in libraries plans that a search can happen purposefully only in the information taken up in the (online) catalogues of the library. These informations are provided according to an appropriate set of rules and is stored in a data base system. The search can access only the text, which is in the title, and/or appropriate contents-opening components such as classifications.

With resources electronically available frequently a full text indexation is offered. However in the decade of the Internet it has the consequence that the hit rate of such systems is extremely high and the relevance of the found information for the user becomes ever smaller. Newer standardisations within the range of the meta data retrieval of electronic resources, as the development of the Dublin core set[4], give the possibilities to look for further meta information in electronic documents and to bring these information itself into the documents, which are then usable over the WWW for search machines. A WWW based search in a DC meta data of the HTML documents, because the meta data are currently used only there, limit the search area and the hit rate already purposefully.

Full texts however contain further components, which, as the meta data of the first page, describe the contents. These elements must be marked by the author himself intellectually by semantic and structure-describing tags: as headings, tables, indices, and bibliographies.

Protection of the authenticity and integrity of the data

The protection of electronic documents can be regarded under the following criteria, as they were among other things already represented in [Ohst 1998b].

Physical preservation

A substantial task of a document server is the protection of the physical integrity of the documents. For this a detailed backup and archiving concept belongs. For the preservation of the documents different media (e.g. CD-ROM, MO) are to be used. As the documents on the server must be protected against illegal manipulation, also the backups must become secured.

[4] http://www.oclc.org/oclc/research/projects/core/index.htm

Access protection

The access protection of digital documents represents the basis for further safeguard concepts. So there can be a limited access to the document server to a certain user circle. This can be achieved by relatively simple procedures, e.g. via the use of IP filters or by the assignment of access accounts and passwords. Technical solutions of this safety problems in the Internet are digital certificates, see [Geschonneck 1998], [HU-CA], and special security protocols such as SSL, see [SSL Specification].

Authenticity and integrity

In order to be able to ensure a durable archiving, an protection of the integrity of the document server and the deposited documents is indispensable. While a conventional publication ensures a relatively good protection by the adjustment on the medium paper digital documents can be copied more easily or changed substantially. With electronic theses author, contents and publication time of the documents must be protected conclusive against falsification and doubts about the authenticity. In addition cryptographic concepts as digital signatures (see in addition [Fox 1997], [to Welsh 1991]) and water-marks, see [ACM Security Workshop 1998] are to be applied. By these characteristics integrated into the documents the origin of the documents at each time can be reconstructed.

License management

For some documents it can be meaningful to limit the number of simultaneous accesses e.g. for copyright reasons. For this the use of a license management is necessary. Also the collection of fees for reading from full texts is perspectively surely conceivable. Here procedures are to be used, which permit a detailed account. Paragraph or access protection by page could support certain sales concepts, then a user e.g. free reading of the first chapter of a document could be permitted, while further chapters require the payment of a fee.

Result

Within the range of the digital university publications, particularly electronic theses, new tasks come both to libraries and to computing centres. It is an area, which makes changes in the scope of duties very clear and clarifies the necessity for a change in the past work of these two facilities. Further one can read in [Schirmbacher 1998a] that there are in Germany several initiatives of these service facilities taking the altered tasks into account.

Literature

ACM Security-Workshop 1998: Multimedia and Security: Workshop at ACM Multimedia, Bristol, UK, 12. Sept. 1998, http://www.darmstadt.gmd.de/mobile/acm98/acm_work/index.html

Behme/Mintert 1998: Henning Behme, Stefan Mintert: XML in der Praxis, Bonn: Addison-Wesley-Longmann 1998.

DiML-Dokumentation1.0: Schulz, Matthias: Dissertation Markup Language (DiML) - Archivierungs- und Rechercheformat für Dissertationen nach dem SGML-Standard - Dokumentation der Dokumenttypdefinition. 1. Auflage Berlin, Januar 1999, Humboldt-Universität zu Berlin, http://dochost.rz.hu-berlin.de/epdiss/software/dimldoc.pdf .

Fox 1997: Fox, D.: Beweismittel - Unterschriften auf der Datenautobahn. iX (1997), H. 12, S. 98-100.

Geschonneck 1998: Geschonneck, Alexander: Vertrauen gegen Vertrauen- Die Zertifizierungsinstanz der Humboldt-Universität zu Berlin. RZ-Mitteilungen Nr. 16 / Juni 1998, Humboldt-Universität zu Berlin, Rechenzentrum, http://www.hu-berlin.de/rz/rzmit/rzm16/2.html .

Goldfarb 1990: Goldfarb, Charles F. The SGML Handbook, Oxford, 1990.

HU-CA: Die Zertifizierungsinstanz der Humboldt-Universität, http://ca.hu-berlin.de/

Kipp 1999: Kipp, Neil A.: Beyond the Paper Paradigm: XML and the Case for Markup, in: In Part II, "Guidelines for Writing and Designing ETD's," ETD Sourcebook, Weisser, Moxley, and Fox, editors., DRAFT: March 17, 1999, http://csgrad.cs.vt.edu/~nkipp/etdsb/ .

Martin 1999: Martin, Norbert: Und wie kommt die Dissertation auf den Server? - Gedanken zum Workflow - Vortrag auf dem Workshop "Workflow", Tagung "Dissertationen Online", 24.03.1999 in Jena, http://dochost.rz.hu-berlin.de/epdiss/jena3/workflow.html .

Meinhold, Luckhard 1998: Meinhold, M. und Luckhard, N.: Echtheits-Zertifikat - Digitale Signaturen mit beweiskräftigem Zeitstempel. c't (1998), H. 8, S. 112-116.

Ohst 1998a: Ohst, Daniel: Dateiformate für das elektronische Publizieren, Studienarbeit am Institut für Informatik, Humboldt-Universität zu Berlin, Berlin, 1998, http://dochost.rz.hu-berlin.de/docserv/buecher/ohst-daniel/HTML/ .

Ohst 1998b: Ohst, Daniel: Was ist ein Dokumentenserver?; Humboldt-Universität zu Berlin: Vortrag auf dem Kolloquium des Rechenzentrums am 10.06.1998; http://dochost.rz.hu-berlin.de/epdiss/kolloqu/ohst/ohst.html.

Rieger 1995: Rieger, W.: SGML für die Praxis. Berlin u.a.: Springer-Verlag 1995.

Schirmbacher 1998a: Schirmbacher, Peter: Die elektronischen Publikationen als Beispiel der Zusammenarbeit von Bibliothek und Rechenzentrums, Vortrag auf dem Kolloquium des Rechenzentrums am 10.06.1998; http://dochost.rz.hu-berlin.de/epdiss/kolloqu/schi/index.htm .

Schirmbacher 1998b: 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 .

SSL-Specification: SSL 3.0 SPECIFICATION, http://home.netscape.com/eng/ssl3/

Welsh 1991: Welsh, Dominic: Codes und Kryptographie. Weinheim, New York, Basel, Cambridge, VCH 1991.