Thinking the long term: the XML-based publishing Workflow for handling electronic theses and dissertations at Humboldt-University Berlin

Susanne Dobratz
Head Electronic Publishing Group of Computing Centre/University Library,
Humboldt-University Berlin, Germany

Keywords: Repository Certification, XML-based electronic publishing, Open Access Publishing, Long Term Preservation

ABSTRACT

This paper gives an overview on how digital theses and dissertations are handled within a cooperative group of the University Library and the Computer- and Media Services at Humboldt- University against the background of a taking over the long term preservation responsibility, including the idea of becoming a trusted digital repository and using an XML- based publishing philosophy.

1. INTRODUCTION

As the emergence of electronic publications grows, it changes the responsibility and duties of the University's Libraries as well as the Computing and Media Departments. Universities acting as publishing bodies are responsible for their digitally generated scientific output and have to guarantee a long term access to e.g. digital dissertations. They can either fulfil this duty through cooperating with National Libraries and Archival Institutions or establish an own so called "Trusted Digital Repository", with all requirements and difficulties.

Humboldt-University started setting up *edoc-Server*, the Document and Publication Repository (http://www.edoc.hu-berlin.de) in 1997. From the very beginning the development focused on two main tasks:

- **Open Access**, which means to provide access to publications, that are like electronic theses and dissertations (ETDs) either usually buried in libraries or are not freely accessible by a major part of the scientific community, like journal articles or older scientific literature, the so called rare books.
- **Long Term Preservation**, in order to ensure that the digital documents published via the edoc-Server can still be used by scientists after a longer period of time. This is a questions of readability and document formats as well as a question of ensuring the authenticity and integrity of the publication's content and the authorship and publication date.

Having started as a pilot project to test the technology, the edoc-Server has developed to a standard service maintained by the University Library and the Computer and Media Services meanwhile. It serves 2500 documents, of which approx. 1000 are ETDs, 275 are professorial dissertations, approx. 30 master theses, 6 digitized documents, approx. 180 Public Readings of the university, 4 conferences with 207 papers, 25 documents from series, and special series containing approx. 300 articles as well as hosting 3 e-journals.

The overall goal of the edoc-Server is to become **the** digital publication platform for Humboldt-University's scientists, that first offers the service "electronic publishing" to all members of the university and secondly provides an "Institutional Repository" for „Self-Archiving". To achieve this goal, the Electronic Publishing Group (EPUB) provides on one hand support and consultation regarding technology, intellectual property issues, etc. to potential authors and editors. On the other hand EPUB develops more and more services connected with the provision and usage of digital documents, such as a print on demand service (Proprint) or a digital university press. The *edoc-Server* is part of the university's information infrastructure and provides interfaces to the media portal (Mneme), the learning management system (Moodle) and the digital library (Metalib).

2. Trustworthy digital repositories

There are two main roads to ensure that documents within an ETD repository will be useable over a long period of time, meaning more than 30 years: First, the development of an so called "Trusted Digital Repository" (RLG2002), (CCSDS2002), and second the strategic alliance with an archival institution, e.g. national or subject libraries.

The pure provision of digital documents can be realised easily by implementing an upload interface for authors, where they can deposit their documents and associated metadata on the server.

Giving guarantees on a future readability and usability of the documents however is a highly sophisticated task. Such guarantees depend on organisational and technical parameters, a document server has to fulfil, e.g.:

- the hard- and software environment,
- the document format, in which the publication is stored,
- the quality and quantity of metadata for that document,
- the quality of securing authenticity and integrity of the archived documents as well as protecting the digital repository itself against misuse, and
- the organisational and financial plans for the maintenance of the repository,
- the transparency of used technology, tools and procedures including their documentation

In Germany a first step has been taken to establish a certain technological level for the document servers, the DINI Certificate for document and publication repositories (DINI2003). Humboldt-University's edoc-Server has reached this certificate and consists therefore of the following features.

3 Organisational features of the *edoc-Server*

Policy of the edoc-Server

With the edoc-Server policy the Library and the Computer and Media Service of Humboldt-University state as institutions, that they will guarantee a reliable and organisationally assured operation of the server. They took over this responsibility by delegating permanent staff into a joint "Electronic Publishing Working Group" (EPUB) and integrating electronic publishing into their common services.

The responsibility of the library is formulated as follows:
"The collection mandate of Humboldt University Library consists of collecting, cataloguing, and archiving all the scientific documents published by the members of Humboldt University. It refers to digitally born documents as well as digital versions of printed documents.
Also included are significant historic documents from the University Library and other institutions that are digitised due to terms of content, conservatory aspects, or the requirements of place-independent use."

The policy also defines the objectives and criteria for the content of the edoc-Server as well as the criteria for handling digital documents. One essential point made in the policy is that once a digital publications has been issued through the edoc-Server it cannot be revoked. Instead a new version of the document could be published. This important statement distinguishes the edoc-Server from e.g. the learning platform or the media portal, where not only permanent digital objects are stored, but also a temporarily storage and presentation is possible.

Legal issues

The legal basis for publishing ETDs in Germany is the law, that for receiving the final degree a publication of the dissertation has to be done. The electronic publication on the university's server has become a common way to fulfil this duty. Other ways are to produce a certain amount of paper copies for the university library or to place a publication in a well-known journal or publishing house. In order to be able to offer additional services in connection with digital publications, the authors have to sign a specific author publishing contract with the university library, where the university has reserved the non exclusive copyright.

So far all documents are Open Access and freely available on the edoc-Server.

Authors support

In order to receive convertible original we have programmed a selection of author templates for different word processing systems, such as Microsoft Office, Staroffice, Openoffice, LaTeX. To ensure a correct usage of those templates we offer consultancy services and support via web pages; e-mail, telephone. The aim is to support the entire publication process and to reach the authors before they start their actual work on the ETD. So we offer specialized courses on " structured writing" for authors every month.

Maintenance of the service

The edoc-server and the services are basically maintained by the EPUB Group. The technical maintenance of the hardware and the operating system is done by experts from the Computer and Media Service, whereas the formal and subject cataloguing is done by special librarians.

Usage of adequate technology and Transparency of service, technology and workflow

In order to become a trustworthy digital repository it is essential to review and on occasion renew the used technology. EPUB does this by a following a consequent project acquisition policy. This means as the university cannot finance all necessary technological developments by itself and in order to be able to perform a technology watch and to maintain cooperation with other institutions, e.g. NDLTD, EPUB produces new project and innovation proposals to funding bodies for the edoc-Server every year.

In order to guarantee transparency of the used procedures and technology, a demand essential for being a trusted and reliable digital repository, a complete documentation of all processes and features is produced permanently and a web server statistic is generated using Webalizer.

4 Technological features of the *edoc-Server*

Addressability

Ensuring reacheability of digital documents and a stable citation independently from URLs, which often change over time, the edoc-Server uses *Uniform Resource Names (URN)* in Form of the *National Bibliographic Number (NBN)*. This is a structured expression containing the kind of urn, the country, the library consortium and the publishing body responsible for the digital publication. In addition a unique identifier for that particular document including a check number is included. The URN-Gateway[1] maintained by the German National Library (DDB) and the Browser-Plugins developed within the *EPICUR-Project* allow to access ETDs and other digital documents via a common web browser. The browser than interprets a URN put into the address line and redirects to the resolving server of the DDB.

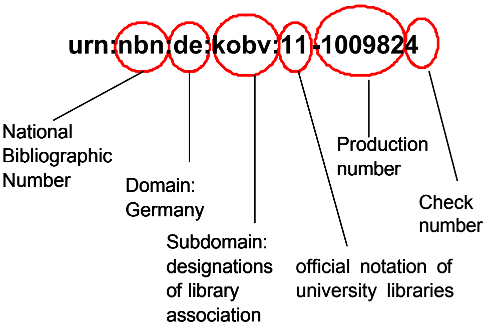
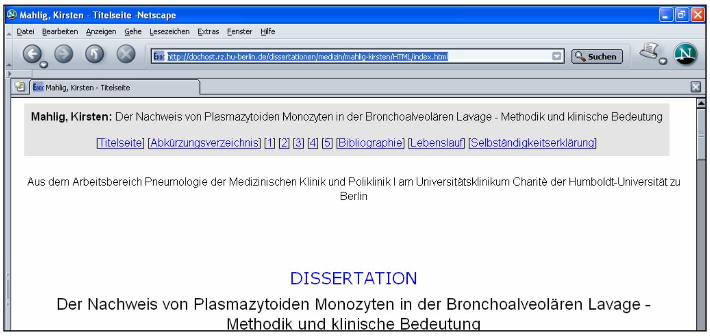

urn:nbn:de:kobv:11-1009824
National Bibliographic Number
Domain: Germany
Subdomain: designations of library association
official notation of university libraries
Production number
Check number

figure 1: composition model of the URNs used for the edoc-Server


Mahlig, Kirsten - Titelseite -Netscape
Datei Bearbeiten Anzeigen Gehe Lesezeichen Extras Fenster Hilfe
Suchen
Mahlig, Kirsten - Titelseite

http://dochost.rz.hu-berlin.de/dissertationen/medizin/mahlig-kirsten/HTML/index.html

Mahlig, Kirsten: Der Nachweis von Plasmazytoiden Monozyten in der Bronchoalveolären Lavage - Methodik und klinische Bedeutung

[Titelseite] [Abkürzungsverzeichnis] [1] [2] [3] [4] [5] [Bibliographie] [Lebenslauf] [Selbständigkeitserklärung]

Aus dem Arbeitsbereich Pneumologie der Medizinischen Klinik und Poliklinik I am Universitätsklinikum Charitè der Humboldt-Universität zu Berlin

DISSERTATION

Der Nachweis von Plasmazytoiden Monozyten in der Bronchoalveolären Lavage - Methodik und klinische Bedeutung

figure 2: urn:nbn:de:kobv:11-1009824

Providing such a service for German publishers pushes the usage of persistent identifier in form of a URN also for ETD repositories in Germany. Humboldt-University is one of 87[2] institutions in Germany that are already using URNs.

Securing integrity and authenticity of the edoc-Server

In order to prevent misuse of the server, the hardware itself is deposited in a special computer room with restricted access that can only be entered by authorized personnel. The same security measures apply for the network connectivity of the edoc-Server, only authorized users can administrate the server and the documents. Secure Shells are used and permanent system maintenance is provided by specialist of the Computer- and Media Service. Staff members in charge of operating the server and the ETD service have certain roles connected with certain permissions, e.g. there are the roles: thesis checker for Word, thesis checker for LaTeX, system manager, a person, that applies the digital signatures and a cataloguer role. This roles also apply for the workflow system that is in use for transferring documents and metadata to the actual server.

Also a backup is run for the server and the edoc-Server is integrated into the distributed storage system (Tivoli Storage Management by IBM) the Computer- and Media Services runs for the whole university at 3 locations.

Securing integrity and authenticity of the documents

In order to safeguard intellectual property rights Humboldt-University uses digital signatures according to the German Signature Law. The reason is to guarantee at least a minimum level of authenticity and integrity of the documents. Through digital signatures one can easily prove if a digital document on the server has illegally been changed since it has been published on the *edoc-Server*. One can also state the exact time of the publication in case an important scientific discovery or a patent has to be proved.

[1] http://www.persistent-identifier.de/
[2] As for 1 July 2005

Digital documents are signed by the following procedure: first a digital hash code is calculated by a cryptographic hash algorithm and than a digital signature for that hash code is generated under use of a personalized digital certificate. If a user wants to check whether a digital document has been altered or not, he has to produce a new hash value that can be used to prove if bits within the signed document have been altered. This is due to the capability of the hash algorithm, that produces a completely different hash code if only one bit in a document has changed. The digital signatures above the hash code states, that a trusted person (e.g. a staff member holding a signature card with a personalized digital certificate) has used his digital certificate at a certain time to state that a particular document or the particular hash code had existed at that time. For the edoc-server we use a service by the German Telekom[3], that runs an own legally approved certification authority and includes the use of smartcards holding the digital signatures and a special cryptographic software. They also maintain a registry for the signatures, so potential reading users could prove the authenticity of the digital signatures on the ETDs.

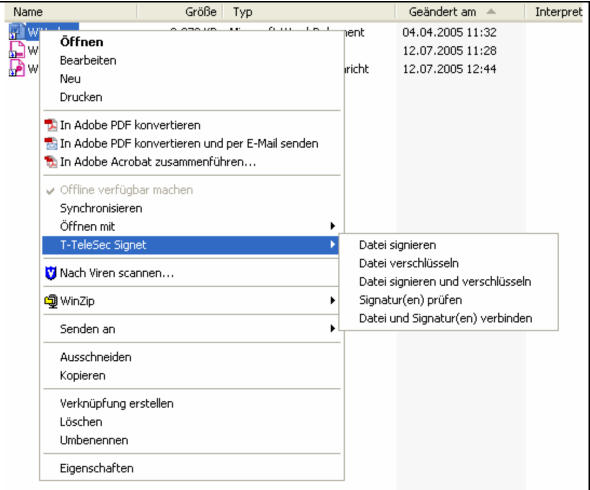
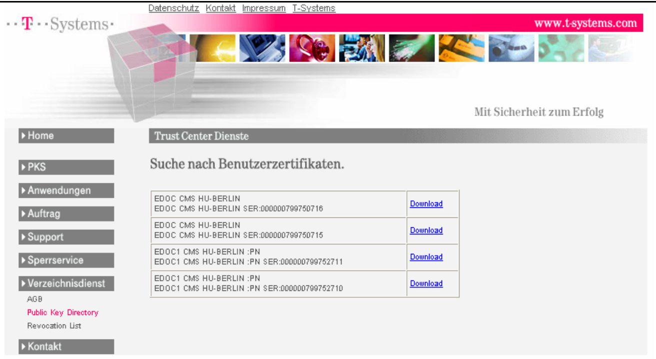


figure 3: Signature software in use



Trust Center Dienste

Suche nach Benutzerzertifikaten.

EDOC CMS HU-BERLIN EDOC CMS HU-BERLIN SER:000000799750716	Download
EDOC CMS HU-BERLIN EDOC CMS HU-BERLIN SER:000000799750715	Download
EDOC1 CMS HU-BERLIN :PN EDOC1 CMS HU-BERLIN :PN SER:000000799752711	Download
EDOC1 CMS HU-BERLIN :PN EDOC1 CMS HU-BERLIN :PN SER:000000799752710	Download

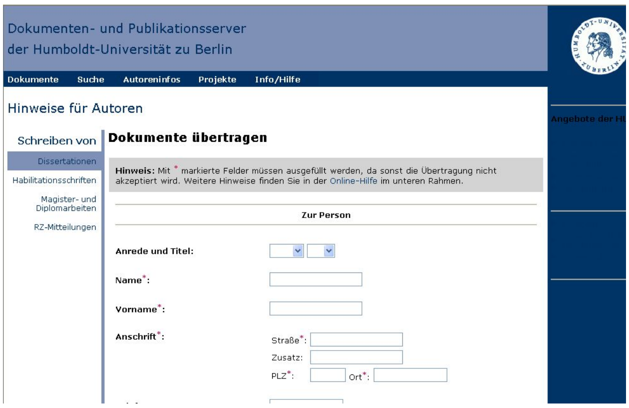
figure 4: digital certificates used for the edoc-Server displayed in teh Public Key Registry of the Telesec Service

Cataloguing

For the cataloguing a special metadata database with a dynamic concept is in use for the edoc-Server. Here the librarian with the role “cataloguer” enters bibliographic metadata, digital signatures, and other additional metadata. For every ETD a German and an English Abstract as well as four keywords each are required from the author. Every ETDs receives a notation from the Regensburger Verbundklassifikation (RVK)[4] and a DDC Subject Heading by the subject librarian in charge of the field the ETD is assigned to.

[3] http://www.telesec.de

[4] RVK is a special subject cataloguing system used in Germany comparable with the LOC subject headings.



Hinweise für Autoren

Dokumente übertragen

Hinweis: Mit * markierte Felder müssen ausgefüllt werden, da sonst die Übertragung nicht akzeptiert wird. Weitere Hinweise finden Sie in der Online-Hilfe im unteren Rahmen.

Zur Person

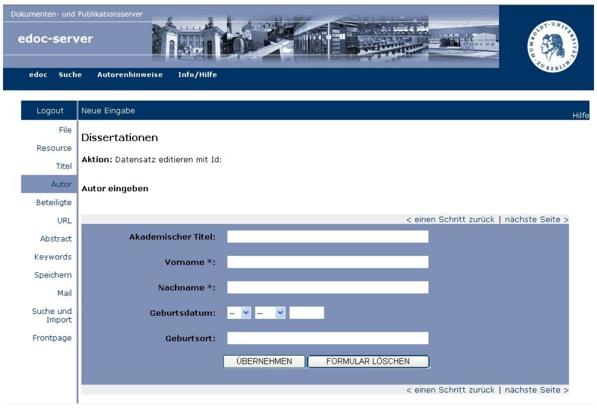
Anrede und Titel:

Name*:

Vorname*:

Anschrift*: Straße*: Zusatz: PLZ*: Ort*:

figure 5: Screenshots of the upload interface provided for authors



Dissertationen

Aktion: Datensatz editieren mit Id:

Autor eingeben

Akademischer Titel:

Vorname *:

Nachname *:

Geburtsdatum: -- --

Geburtsort:

figure 6: Screenshots of the metadata database user interface

Some of the technological metadata , such as URNs or location on the server is generated automatically. Tools like JHOVE7, that enable an even more automated generation of technical metadata about document formats, versions etc. are being tested by the edoc-Server staff and will be integrated into the whole concept within future projects.

Workflow Management

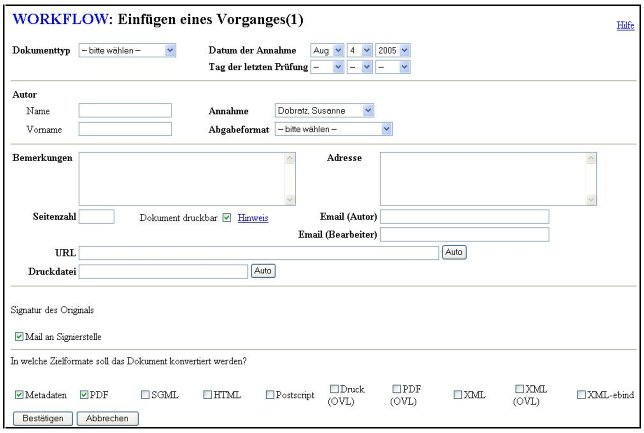
In order to support different roles for staff members that carry out different types of work, a workflow system was implemented quite early. It provides support and information for the staff members, not for the authors or publishers. The last issue on one that is actually been exploited by the SCOPE[5] project.

It was designed to support the following ETD-Workflow:

1. Receive an upload done by an author
2. Check completeness and technical correctness of the ETD
3. Accept ETD for publication on the edoc-Server
4. Inform author about acceptance or non-acceptance
5. Inform theses librarian about acceptance. He will than send out a receipt to the faculty which the author needs to rece the grade and the graduation certificate.
6. Initiate digital signature process for the received original (mostly word or LaTeX documents) and the PDF version of the ETD, that is used for producing printed copies
7. Initiate conversion into archiving format XML
8. Initiate cataloguing
9. Initiate uploading of accepted and converted versions of the ETD to the edoc-Server and giving it free for public access.
10. Inform author about each separate publication step
11. Initiate the printout of hard copies and state the receipt of the print charge necessary for completing the publication.

Initiating further steps in the workflow is done by sending Email to staff members that are associated with a particular role. For the conversion process a period of two weeks is automatically generated, after which the light system changes the light from green to red, which indicates for the processing person that the importance to solve that issue has become more urgent.

[5] See http://edoc.hu-berlin.de/scope/



WORKFLOW: Einfügen eines Vorganges(1)

Dokumenttyp: – bitte wählen –

Datum der Annahme: Aug 4 2005

Tag der letzten Prüfung: – – –

Autor

Name:

Vorname:

Annahme: Dobratz, Susanne

Abgabeformat: – bitte wählen –

Bemerkungen:

Adresse:

Seitenzahl:

Dokument druckbar ☑ Hinweis

Email (Autor):

Email (Bearbeiter):

URL:

Druckdatei:

Signatur des Originals

☑ Mail an Signierstelle

In welche Zielformate soll das Dokument konvertiert werden?

☑ Metadaten ☑ PDF ☐ SGML ☐ HTML ☐ Postscript ☐ Druck (OVL) ☐ PDF (OVL) ☐ XML ☐ XML (OVL) ☐ XML-ebind

figure 7: Screenshots of the workflow database user interface

5 An XML- based publishing approach

Nevertheless, an important part within a long term preservation environments is to ensure that the “Ingest” part of the system is fed with documents that have a certain quality technically and regarding the intellectual content. At Humboldt-University we have therefore decided in 1997 to follow an XML- based publishing approach and formulated a server policy. The strategy focuses on the conversion of documents being delivered by the authors in standard word processing formats like Microsoft Word, Staroffice / Openoffice into an XML- format according to the xDiML Document Type Definition. The experiences and tools were also taken for evaluation and distribution within the Germanwide Dissertation Online project, which was funded by the Deutsche Forschungsgemeinschaft from April 1998 until October 2000. This technology is steadily improved within the XML based publishing platform.

The XML based approach is performed in 2 ways:

1. an author-based approach, where the conversion from a word processing file into XML is done as a service by the EPUB Group. This applies for dissertations, professorials dissertations, digitized materials
2. the editor-based approach used the SCOPE platform and provides tools for editors and publishers and they are made responsible for the conversion. They deliver the XML documents to the edoc-Server.

Why XML for digital publication?

Answering the question which the document format suits best to guarantee a long term preservation, the following points seem to be essential to be considered. The ideal solution would conserve the full originality of a digital object, making it as easy to use emulation strategies or just to bring that object to function within a new hard and software environment. This would preserve the original look and feel. Often this idealiter cannot be reached, so compromises have to bmade that take a loss of functionality or features into account. Here it is most important, that the characteristic features of an object can be preserved. Therefore following criteria are useful in order to evaluate the long term preservation ability of a format, see also (Stanescu2004):

1. documentation and publication of the format
2. standardisation of the format by an official body
3. modularity of the format, e.g. separate between markup, layout and content of a document (like XML and other markup languages do)
4. functionality of the format
5. distribution of the format, costs for using it, existence of tools
6. special licensing conditions or costs

1. availability of the source code
2. documentation of revisions, backwards compatibility

3. quality of format documentation in general
4. is the format widely used or just within small communities?
5. are there similar and concurrent formats?
6. can Digital Rights Management systems, metadata or digital signatures be used?
7. are there experts fort his format?
8. kind of revision cycles
9. additions and spezial features
10. can the authenticity of a document be altered easily? How good apply technologies like digital watermarks?
11. is the community associated with the format big enough to conserve the format (e.g.LaTeX)

For Humboldt-University the decision was made for SGML and XML[6] because markup languages can fulfil the demand for a long term preservation format best, although they also have some disadvantages, lying in the media capability, in the standardisation of specifically used formats e.g. for ETDs and in their lack of processing efficiency.

6 Long term preservation and Open Access

We have often discussed the XML philosophy which aims towards long term preservation against the background of realising an easy to use Open Access component also for scientific articles previously published in scientific journals. Here the pure provision of the documents has a become a higher priority than the long term preservation argument and we decided to offer additional services for conversion but on the other hand provide an Open Access Interface, that allows a simple upload for documents. The only restriction we make is to accept only certain document formats, like PDF, where we try to promote the PDF-X or PDF-A usage and HTML, (focusing on XHTML) because this is a basic XML bound format.

7 REFERENCES

DINI2003: DINI Certificate Document and Publication Repositories, Deutsche Initiative für Netzwerkinformation, 2003, http://www.dini.de/documents/Zertifikat-en.pdf

EDOC2002: Document and Publication Server of Humboldt University Berlin, - Policy -, 2002, http://edoc.hu-berlin.de/e_info_en/policy.php

RLG2002: RLG/OCLC Working Group on Digital Archive Attributes: *Trusted Digital Repositories: Attributes and Responsibilities*. Mountain View, CA, RLG, 2002, www.rlg.org/en/pdfs/repositories.pdf(28.04.2005)

CCCSDS2002: Consultative Committee for Space Data Systems (CCSDS): *Reference Model for an Open Archival Information System (OAIS)*, CCSDS 650.0-B-1, BLUE BOOK, 2002. www.ccsds.org/documents/-650x0b1.pdf (28.04.2005).

Stanescu2004: A. Stanescu: Assessing the Durability of Formats in a Digital Preservation Environment, D-Lib Magazine, Nov. 2004, (vol 10, no 11), http://www.dlib.org/dlib/november04/stanescu/11stanescu.html

[6] In 1997 we started with SGML and when XML appeared in 1998 we changed to that concept