Long-term preservation of Electronic Theses and Dissertations

Hans Liegmann
Die Deutsche Bibliothek
liegmann@dbf.ddb.de
Adickesallee 1, D-60322 Frankfurt am Main, Germany
http://www.ddb.de
Keywords: long-term preservation, deposit SYSTEM for electronic publications, Open Archival Information System, OAIS, document server, NEDLIB, Die Deutsche Bibliothek

Abstract
Since 1998 online dissertations are collected and archived by Die Deutsche Bibliothek on a voluntary basis. Online dissertations are a special type of electronic publication. The task of long-term preservation of these publications is in principle the same as for electronic publications in general. Based on the process model of the Deposit System for Electronic Publications designed by the NEDLIB project a process has been developed by Die Deutsche Bibliothek. The desired result of this process, which is meant to be never ending, is the long-term availability of all online dissertations collected. This process is described by a rather practical point of view. The stages that have to be passed through until a publication can be accessed are identified and are described. Tools that are deployed in this process are mentioned. The role metadata play in this process is explicated. Statistics of actual use are presented. Furthermore cooperations with other institutions that improve the accessibility are addressed.

Preface

Die Deutsche Bibliothek[1] is the national library and national bibliographic information centre for the Federal Republic of Germany. It is responsible for the collection, processing and bibliographic indexing of all German and German-language publications issued since 1913. Die Deutsche Bibliothek was established in 1990 on the basis of the Treaty of Unification in a merger of the existing institutions the Deutsche Bücherei Leipzig and the Deutsche Bibliothek Frankfurt am Main, of which the Deutsches Musikarchiv Berlin is an integral part. Die Deutsche Bibliothek has a legal mandate to collect, index and permanently preserve the total of German-language publications and publications issued in Germany. The law also includes digital publications as far as they are supplied on physical carriers. Online publications are not yet covered by the legal mandate. For several years Die Deutsche Bibliothek jointly with publishers and producers has been testing the delivery, archiving and long-term preservation of online publications on this basis. In this work group "Electronic Deposit Library" the conditions were tested and negotiated under which Die Deutsche Bibliothek can act as archive also for online publications. Die Deutsche Bibliothek collects Electronic Theses and Dissertations since 1998. Metadata for ETDs are being reported by the university libraries to Die Deutsche Bibliothek in a standardized format. The electronic texts themselves are then actively and individually transferred by our library staff onto our archive server http://deposit.ddb.de. Until now, about 16.000 online dissertations have been collected following this procedure.

ETDs - a good start for preservation activities?

Most of our activities in the field of electronic publications started with ETDs as "candidates". Why? You may assume, it is because we judge them as the core of our digital cultural heritage. Yet, in the first place it is because we prefer to start new workflows using quite simple object categories. ETDs have a quite simple bibliographic appearance: monographic, finite and stable without versioning problems, without complications like "successively issued" and "integrating" publication types. Working on long-term preservation and perpetual access for digital resources is a task of considerable complexity. It needs an encouraging start. The second reason for our choice was, that ETDs publishing process is not driven by the publishing market and by commercial players with their very different interests compared to an archive or a library. Our colleagues at universities and university libraries are cooperative and friendly people. They understand, share and promote our ideas about long-term availability. It is also exciting to talk to commercial publishers about implementation of delivery procedures for digital resources and metadata, but when it comes to assigning manpower, cooperation can get very difficult. The third point relates to the second: our colleagues at the universities can even influence the creation process for ETDs to support "preservation friendliness". Rising awareness, stipulation of publication workflows and document structures including practical support are actual examples for this active influence.

As a national legal deposit library, we have to be conscious about the "real world problems", when we start

1 http://www.ddb.de

building our long-term preservation strategies using ETDs: digital resources in general are very heterogeneous, signs for standardization are still hardly to be seen. Their appearance is determined by the economic conditions of the publication market and the requirements of present-day users. The needs of our children and grandchildren have to be kept in mind and ensured by libraries and archives. It is their key role to preserve and guarantee access to our scientific and cultural heritage as recorded in publications and archival records. The decision of what has to be secured for the future should not be based on technical simplicity and feasibility, but on long-term content value.

Document server and deposit SYSTEM

A deposit SYSTEM for ETDs or for digital resources in general has to concentrate on very specific tasks, building a small and somewhat exotic subset of what we are used to call "digital library" functionality. New self-publishing concepts have taken over archiving terminology and challenge its contents. Arnoud de Kemp from Springer publishing house (Heidelberg, New York, Berlin) said once "Publishing an e-journal is more than just putting it onto a web server". Using his example, I would derive the statement "Archiving is more than just implementing a web server and storing ETDs on it". A deposit SYSTEM is built for the future, and it is sometimes hard to show its benefits and efficiency in the present. Therefore, we follow the concept of strict functional separation: library management SYSTEM, document server and deposit SYSTEM should be interoperably based on open interfaces and protocols, and it should always be possible to cope with strengths and weaknesses of each building block separately. Core business of deposit libraries and mass usage oriented libraries is different: they complement each other concerning high performant access and prodeures for guaranteeing long term avaliability. This also affects responisibilities for ETD life cycle management.

Deposit SYSTEM - the OAIS reference model

A deposit SYSTEM supports procedures to preserve electronic publications (online and offline) and keeps them accessible through time. To implement such a SYSTEM, specific standards, technologies, and procedures are needed. Project NEDLIB - Networked European Deposit Library[2], funded by the European Union from 1998 to 2000, developed a process model for deposit libraries. On the basis of the Reference Model for an Open Archival Information System - OAIS[3], European national libraries and archives found common grounds to encourage applied research in the area of digital preservation. The OAIS generic model consists of functional entities with well-defined interfaces and introduces a concept of packages to standardize and interconnect preservable content and metadata. A digital object is prepared for submission to the deposit SYSTEM and packaged into a SIP (submission information package). The functional entity "Ingest" is responsible for preparing storage, further identification and creation of necessary metadata. Afterwards, repackaging into an AIP (archival information package) takes place. Core components of the model are "Archival Storage", where services and functions for the storage and maintenance of AIPs are provided, and "Preservation Planning". The latter entity monitors the environment of the archival SYSTEM and provides recommendations to ensure that the information once stored remains accessible to the user community over the long-term even if the original computing environment becomes obsolete. When requested for access, the AIP is prepared for delivery and rebuilt as a DIP (dissemination information package). All these packages are more or less still conceptual views, which have to be technically implemented in reality for all kinds of digital object types. Nevertheless, we committed ourselves to OAIS compliance for our actual SYSTEM development strategy.

Submission packages for ETDs

A considerable amount of our capacity in the last years had to be invested into the implementation of transfer procedures for digital resources into our electronic library stacks.

Our experience with ETDs showed that submission information package definition is in principle not too complicated for this type of material. Most of our dissertations and theses consist of one single file (90% are in PDF format) and in opposite to some early predictions, they still do not bring along extensive additional material like rotating molecule models, multimedia additions, executable programmes or data sets. But we wanted to handle the few and rather simple examples for multifile objects in a standardized way to prepare ourselves for future complexity. If consistency and completeness of multifile objects (e. g. a bunch of HTML files) has to be guaranteed, no one does better than the author or primary

2 http://www.konbib.nl/nedlib
3 http://ssdoo.gsfc.nasa.gov/nost/isoas/

publisher. Following this rule, we pragmatically defined a container format for multifile ETDs in 1999. It uses a choice of archive formats (ZIP, TAR) to keep together the dependent parts of the document. Additionally, we introduced a simple table of contents file in order to standardize the root element for future migration activities even on single files and for user access and navigation.

Since 1999, several projects and initiatives including ourselves have worked on proposals for submission information packaging. We are looking at the standardized container format for eBooks[4], Harvard[5] project results on eJournal article transfer and we are observing with great interest the deployment of the Metadata Encoding and Transmission Standard METS[6] . A METS document consists of five components: descriptive metadata, administrative metadata, file groups, structural map and may even include behaviour. METS' ability to combine metadata, content and structural information in one entity makes it very attractive for digital object transfer. METS has its roots in the Digital Library Federation, so that openness is provided for and further input from library and archive communities is possible. We envisage METS as a possible successor for our pragmatic "homegrown" container format for ETDs and other digital objects.

Until now, ETDs archival copies are pulled into the archive by librarians of Die Deutsche Bibliothek individually and manually. Due to the growing amount of ETDs and the sometimes problematic response performance of the original servers, we are interested in automating this procedure as far as possible. We will soon start an experiment, where we will harvest ETD servers using 'Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)'[7], then extract the (preferably persistent) document identifier from the metadata set and finally pull the ETD by a robot automatically.

Metadata for preservation

ETDs life cycle in Germany is accompanied by METADISS[8], a set of qualified Dublin Core metadata elements describing relevant document properties for data exchange between document servers of universities or university libraries and Die Deutsche Bibliothek. METADISS came into being when work about metadata elements for preservation was still in its infancy. Meanwhile several projects, including NEDLIB, have proposed specialized metadata sets of variable granularity in different application scenarios. The OCLC/RLG Working Group on Preservation Metadata has published a synoptic view[9] of the most important sets as a contribution for the cooperative development of a preservation metadata framework. Today, METADISS does not sufficiently reflect the need for preservation metadata and will have to be enriched soon accordingly. In 2001, Die Deutsche Bibliothek has implemented a submission interface for online publications[10], produced by publishers and other institutions. During the submission procedure, we also ask for technical metadata relevant for preservation purposes. We tried to find a compromise between the workload publishers are willing to bear under the conditions of voluntary submission, and the extensive requirements of future preservation processes in our deposit SYSTEM. A moderate solution was the definition of so called "reference SYSTEMs", representing the software and hardware requirements for a certain publication type during a period of time. I. e. we do not investigate too many technical details about those requirements, which are presently customary in the market. Instead, we record extraordinary conditions the publication needs for rendering. Nevertheless, the level of metadata granularity still has to be discussed and, even more important, a consensus between metadata creators, intermediaries and end users has to be found. In our view, National Library of New Zealand's Metadata Standards Framework for Preservation Metadata[11] has a high potential for bridging the gap between practicalities of metadata creation, capture and the high level conceptual principles of a preservation framework. The importance of persistent identifiers as part of a metadata framework for ETDs will be covered by one of my colleagues from the project "EPICUR - Enhancement of Persistent Identifier Services - Comprehensive Method for unequivocal Resource Identification"[12] during another session in this conference.

Trusted repositories: bit preservation and document rendering

The OAIS reference model building block "Archival Storage" is responsible for keeping intact the bitstream of the digital object's preservation master. It may either be the

4 http://www.openebook.org/
5 http://www.diglib.org/preserve/harvardsip10.pdf
6 http://www.loc.gov/standards/mets/
7 http://www.oclc.org/research/pmwg/background.shtm
8 http://www.deposit.ddb/metadiss.htm
9 http://www.oclc.org/research/pmwg/background.shtm
10 http://deposit.ddb.de/netzpub/web_abgabe_np_gesamt_e.htm
11 http://www.natlib.govt.nz/files/4initiatives_metaschema.pdf
12 http://www.persistent-identifier.de/

original bitstream or the result of one of several migration procedures in the archive life cycle of the digital object.

Whatever strategy will be followed in the future to provide access to the digital content, it will depend on the existence of a bitstream, the integrity and authenticity of which has been kept in order over the years. It needs more than a RAID-5 disk storage SYSTEM with redundant backup to guarantee this. Again, an OCLC/RLG working group has done groundbreaking work. The report "Attributes of a Trusted Digital Repository"[13] has articulated a framework of attributes and responsibilities for trusted, reliable and sustainable digital repositories. As bit preservation strategies are well known and well tested in applied information technology, the challenge is rather organisational than technical. The report proposes to use and to formalize certification procedures as a means for proving reliability and trustworthiness of repositories over time. Networked repository services depend on cooperation. Transparency of workflows, definition of service levels and documentation of security provisions is a sound foundation for mutual trust.

On the basis of the preserved bitstream, document rendering will have to be enabled for future access to digital objects. Several strategies are in discussion and a lot of projects are prototyping various technical methods: from migration on request to the concept of a Universal Virtual Computer (UVC)[14] . We assess them all as valuable efforts in order to achieve persistence for digital publications including ETDs.

Our hosts from Humboldt University do important work at the start of the production chain: they invest in promotion of authoring tools and end user services which should provide for well structured and preservation friendly ETDs using SGML/XML formats[15] . All of us presenting papers at this conference have experienced that you need to tolerate and to adhere to a structured framework in order to profit from its advantages.

Our colleagues from the Royal Libray of the Netherlands in The Hague have inaugurated the first large scale implementation of a Digital Information Archiving System[16] . Time is too short to list all the important efforts in Australia, the US and Europe, but you may use "PADI - the Subject Gateway to Digital Preservation Resources"[17] for further information on this topic.

Synergies and possibilities for cooperation will have to be exploited to a maximum extent in order to be able to solve our future problems. It has been a good start using ETDs.

13 http://www.rlg.org/longterm/repositories.pdf
14 http://www.kb.nl/kb/ict/dea/ltp/reports/4-uvc.pdf
15 http://www.edoc.hu-berlin.de/index_en.php
16 http://www-5.ibm.com/nl/dias
17 http://www.nla.gov.au/padi/