Interoperability: new challenges and solutions

28th Library System Seminar

Trondheim (Norway), 9 - 11 June 2004

Workshop: Archiving for digital resources

Reinhard Altenhöner, Die Deutsche Bibliothek, Frankfurt/M., Germany

Electronic publishing is the common form of scientific communication and publication and is becoming increasingly importance for the world of publication in general: easy to handle, worldwide available and accessible at every time. But otherwise digital information objects are in the deep sense of the word much more fragile than traditional paper based information. They can be easily altered, damaged or become unreadable especially in the long run of time. In this situation a special deposit system is necessary, which supports procedures to preserve electronic publications (online and offline) and keeps them accessible through time. In a typical librarian perspective of decenniums there are a lot of questions and unsolved problems: Implementing such a system requires specific standards, technologies, and procedures.

The Aim of the workshop is

- To gain a broad introduction into the problems
- To learn from and discuss the main issues and the strategies for long-term preservation in the worldwide dimension
- To sort and assess practical steps
- To exchange/share experiences on the field of long-term preservation
- To discuss the consequences in the field of organisation, workflow and structures for libraries

The workshop will - without loosing the perspective of international research - be coined by the experiences, the practical implications and solutions of Die Deutsche Bibliothek, the German national library.

Die Deutsche Bibliothek, Frankfurt a. M./ Leipzig/Berlin

Die Deutsche Bibliothek has a statutory mandate for the collection, bibliographic processing and long-term preservation of all publications released in Germany or published in the German language abroad. The law also covers digital publications distributed on physical carriers but makes no provision for online publications. A number of fundamental principles applicable to the collection of online publications were defined in preliminary hearings with publishers, library experts, information specialists and government officials and formulated in a policy document passed by the Publishers' Committee of the Börsenverein des Deutschen Buchhandels in June 1997:

- All online publications are to be submitted via data networks or on physical data media upon request.
- Online publications available in different forms are to be submitted in the format requested by the library.
- Publications with identical contents distributed both on physical media and as online publications are to be submitted in both forms.
- Online publications with identical contents distributed simultaneously by multiple providers need only to be submitted once.

Die Deutsche Bibliothek shall be authorised to produce a copy of each digital publication for the purpose of long-term preservation. Authenticity of publication content must be ensured.

On the basis of these policy principles, Die Deutsche Bibliothek has tested procedures for the submission, collection and long-term preservation of online publications in co-operation with publishers and producers in a test phase lasting several years. In the process, the 'Electronic Deposit Library' task force explored and established the conditions necessary for Die Deutsche Bibliothek to become a deposit library for online publications as well.

The DDB's experiences in the field of digital preservation:

- Dissertations and Theses Online (DissOnline) for archiving dissertations (in practical use) CARMEN-AP4 and EPICUR about implementation and usage of persistent identifiers and co-operation with Springer-Verlag (Heidelberg, Berlin) in archiving eJournals.
- The DDB has built up experiences by their System for Multimedia Access - Multimedia-Bereitstellungssystem (MMB). MMB enables storage and access for digital objects on physical carriers. Different object types (workstation image, application installation kit, file collection, presentation object) have been implemented to provide for the rendering of complex digital objects (applications).

The International perspective

The basic standard of a Reference Model for an Open Archival Information System – OAIS describes the situation in principal (http://ssdoo.gsfc.nasa.gov/nost/isoas/). 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 (the more recently introduced) "Preservation Planning".

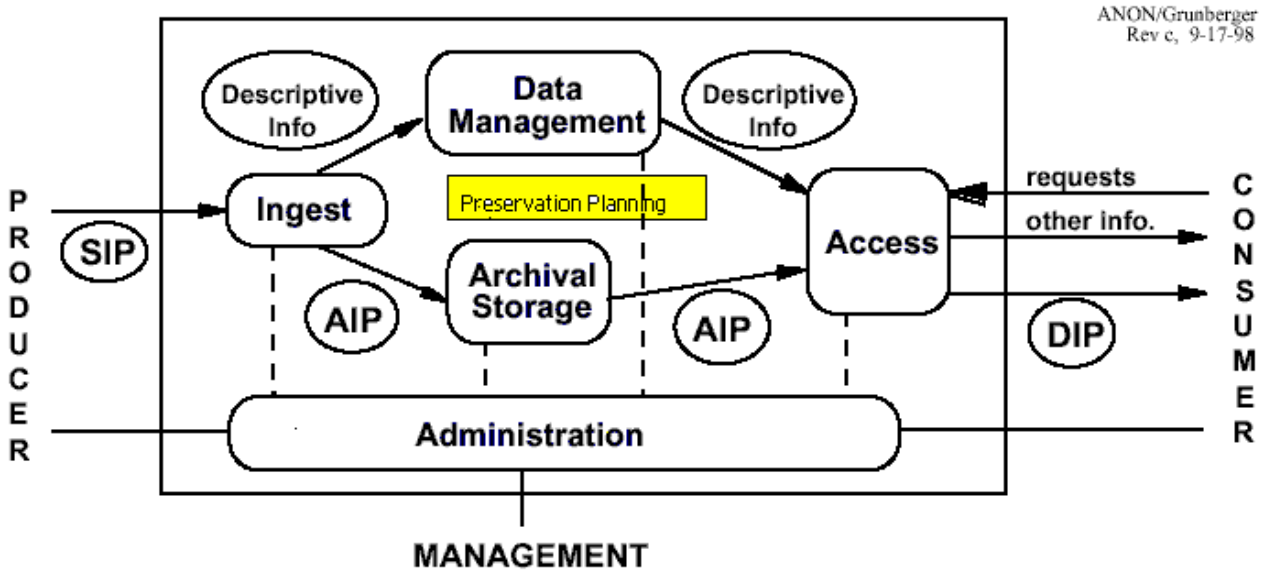


Figure 4-1. OAIS Functional Entities

"Archival Storage" is responsible for keeping intact the bitstream of the digital object's preservation master. It may either be the original bitstream or the result of one of several migration procedures in the archive life cycle of the digital object. A deposit system supports procedures to preserve electronic publications (online and offline) and keeps them accessible through time.

The “Preservation Planning” 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.

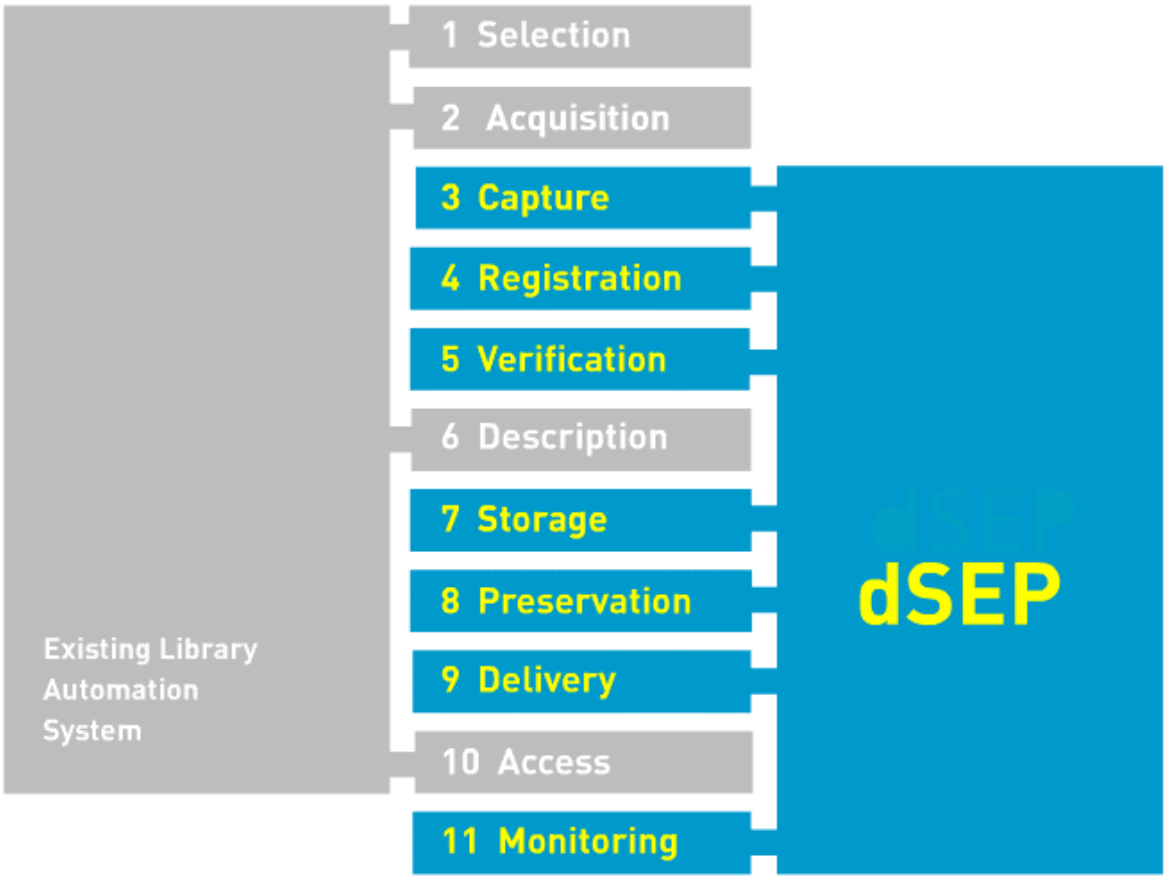
When requested for access, the AIP is prepared for delivery and rebuilt as a DIP (dissemination information package). All these packages are more or less still conceptual views, which have to be technically implemented in reality for all kinds of digital object types. Nevertheless, the most initiatives committed themselves to OAIS compliance for our actual system development strategy.

NEDLIB

The implementation of such a system requires specific standards, technologies, and procedures are needed. The Project NEDLIB - Networked European Deposit Library (http://www.konbib.nl/nedlib), funded by the European Union from 1998 to 2000, developed under participation of DDB a process model for deposit libraries. On the basis of OAIS, European national libraries and archives found common grounds to encourage applied research in the area of digital preservation.

The aim of this project was to define and identify the basic infrastructure-components upon which a networked European deposit library could be built. The objectives of NEDLIB coincided with the mission of national deposit libraries to ensure that current electronic documents could be used now and in the future. The NEDLIB project created a consensus in the way to handle and store digital documents. The digital documents should be separated from its original carrier or environment, which is

intended for publishing and not for archiving, and stored in a controlled archiving environment. Such a controlled environment is currently defined as ‘a safe place’.



The basic principle of the NEDLIB’s process model is to separate archiving from other functions; i.e. searching, authentication and authorisation. The archiving system has to be integrated in the larger ICT-infrastructure of the institution that provides the other functions. This design using separate components ensures durability: every single part can be built and replaced as time goes by without being dependent on one provider and without being too complex.

Trusted repository, Migration & Emulation

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. An OCLC/RLG working group has done groundbreaking work. The report "Attributes of a Trusted Digital Repository" (http://www.rlg.org/longterm/repositories.pdf) 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 of

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 are 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, which can be summarized in two action lines:

1) to migrate the electronic objects in a controlled environment (see above)
2) or to emulate the historic system environment form the origin time of the object – including the emulation of hardware and system software.

A lot of projects are prototyping various technical methods: from migration on request to the concept of a Universal Virtual Computer (UVC) (http://www.kb.nl/kb/ict/dea/ltp/reports/4-uvc.pdf).

Metadata

Since 1999, several projects and initiatives have worked on proposals for submission information packaging. See i.g. the standardized container format for eBooks (http://www.openebook.org/) or Harvard project results on eJournal article transfer (http://www.diglib.org/preserve/harvardsip10.pdf). The deployment of the Metadata Encoding and Transmission Standard METS (http://www.loc.gov/standards/mets/) is of special interest: 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.

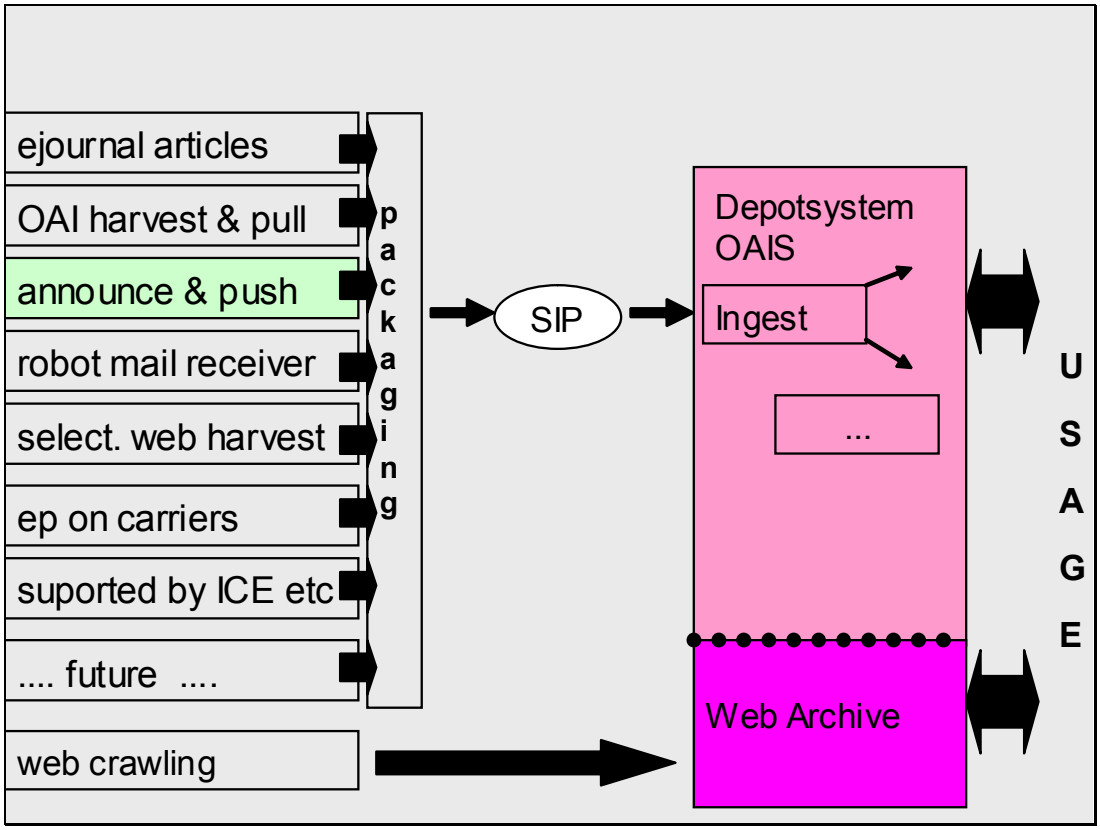
The OCLC/RLG Working Group on Preservation Metadata has published a synoptic view (http://www.oclc.org/research/pmwg/background.shtm) of the most important sets as a contribution for the cooperative development of a preservation metadata framework.

This overview identifies a lot of activities representing the high level conceptual principles of a preservation framework, but the question of practicalities of metadata creation and capture stays out. Perhaps the National Library of New Zealand's Metadata Standards Framework for Preservation Metadata (http://www.natlib.govt.nz/files/4initiatives_metaschema.pdf) bridges the gap.

Persistent identifiers

The importance of persistent identifiers as part of a metadata framework for electronic publications will be covered by another DDB-project "EPICUR - Enhancement of Persistent Identifier Services - Comprehensive Method for unequivocal Resource Identification"(http://www.persistent-identifier.de/).

In a range of tasks one can consider a lot of different activities, which culminate in a scheme. This shows the organizational needs for a deposit system.



In practice: ETDs (Electronic Theses and dissertations)

The experience with ETDs showed that submission information package definition is in principle not too complicated for this type of material. Most of the dissertations and theses in DDB (until now round about 24.000 items) consist of one single file (90 % are in PDF format) and in contrast to some early predictions, they still do not bring along extensive additional material like rotating molecule models, multimedia additions, executable programmes or data sets. 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 publisher. Following this rule, there is a pragmatic definition: a container format for multifile ETDs. It uses a choice of archive formats (ZIP, TAR) to keep together the dependent parts of the document. Additionally, there is a simple table of contents file to standardize the root element for future migration activities even for single file documents including user access and navigation.

Online Publications

In 2001, Die Deutsche Bibliothek has implemented a submission interface for online publications (http://deposit.ddb.de/netzpub/web_abgabe_np_gesamt_e.htm), produced by publishers and other institutions. During the submission procedure, DDB is also asking for technical metadata relevant for preservation purposes. This has to be 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 the 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. there is no investigation on technical details about those requirements, which are presently

customary in the market. Instead, DDB records extraordinary conditions the publication needs for rendering.

For further information

It's About Time: Research Challenges in Digital Archiving and Long-term Preservation
URL: **http://www.digitalpreservation.gov/index.php?nav=3&subnav=11**

The State of Digital Preservation: An International Perspective
Washington, D.C.: Council on Library and Information Rescources, July 2002. ISBN 1-887334-92-0
URL: **http://www.clir.org/pubs/abstract/pub107abst.html**

Neil Beagrie:
National Digital Preservation Initiatives: An Overview of Developments in Australia, France, the Netherlands, and the United Kingdom and of Related International Activity
URL: **http://www.clir.org/pubs/reports/pub116/contents.html**

N. Beagrie, M. Jones:
Handbook Digital Preservation
URL: **http://www.dpconline.org/graphics/handbook/index.html**

Margaret Hedstrom und Sheon Montgomery:
Digital Preservation Needs and Requirements in RLG Member Institutions
URL: **http://www.rlg.org/preserv/digpres.html**

Join Information Systems Committee (JISC):
Digital Preservation
URL: **http://www.jisc.ac.uk/dner/preservation/**

National Initiative for a Networked Cultural Heritage (NINCH):
The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials
URL: **http://www.nyu.edu/its/humanities/ninchguide/**

Reference model for an Open Archival Information System(OAIS)
URL: **http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html**

Jeff Rothenberg:
Ensuring the longevity of digital information (pdf)
URL: **http://www.kb.nl/kb/ict/dea/download/dig-info-paper.rothenberg.pdf**

Trusted Digital Repositories: Attributes and Responsibilities
URL: **http://www.rlg.org/longterm/repositories.pdf**