Data for the future:
The German project "Co-operative Development of a Long-term Digital Information Archive" (kopal)

Reinhard Altenhöner
Die Deutsche Bibliothek (DDB), Frankfurt am Main, Germany

Abstract

Purpose: One of the unresolved problems of the global information society is ensuring the long-term accessibility of digital documents. The project kopal tackles this problem head-on: In a three-year project kopal's objective is the practical testing and implementation of a cooperatively created and operated long-term archival system for digital resources.

Design/methodology/approach: The system will be implemented in accordance with international standards for long-term archiving and metadata within the OAIS framework (Open Archival Information System). The project partners, Die Deutsche Bibliothek (DDB), Göttingen State and University Library (SUB Göttingen), IBM Deutschland GmbH and the Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), will establish a cooperatively transferable solution for cultural heritage institutions, as well as for business and industry.

Findings: Within the project, the project partners DDB and SUB Göttingen are developing software for the input and access of data, which will be released under an open-source license.

Research limitations/implications: Long-term preservation methods and strategies will be discussed in general in the paper.

Practical implications: The project will present a stable and reusable platform for additional partners and users, especially for cultural heritage organisations.

Originality/value: The solution is based on DIAS (Digital Information and Archiving System), jointly devised by IBM and the National Library of The Netherlands in The Hague, and it will be adapted to the needs of the project with several extensions. Establishing a collaborative solution for long-term preservation is a milestone in the development of systems for the long-term availability of digital objects.

Keywords: Digital libraries, Archives management, Digital storage, Germany

Paper type: Research paper

Publication process

Today, the production phase of the publication process is completely electronic-based, but increasingly the same is true for the delivery of publications to the end-users, the readers. But we still do not know how digital publications, works of art, image and sound documents, files, primary data and music can be archived so that they will remain permanently readable and thus accessible to all in the future. What we can say at the present time is: Whatever strategy is followed in the future to provide access to digital content, it will depend on the existence of a bit stream, the integrity and authenticity of which has been kept in order over the years and decades. So bitstream-preservation has the potential facility to make digitally stored objects available for a long time. Based on this, document rendering will have to be enabled for access to digital objects in the future. Several strategies are currently under discussion. The most important action points can be summarized in these two bullets:

- From today's perspective, migrating electronic objects in a controlled environment from one format to another, will be more usual and better for long-term access.
- The alternative is to emulate the historic system environment from the production time of the object – including the emulation of hardware and basic system software.

In addition to the availability of the bit stream, we need metadata information, in particular concerning technical information about the stored object, its original technical environment and its storage and migration history. The task of establishing a system and an infrastructure for long-term preservation is complex and demands a large amount of money and personal resources, and we need an academic network and knowledge-sharing – nationally and internationally.

Die Deutsche Bibliothek [1]

Die Deutsche Bibliothek (DDB) is the national library and national bibliographic information center for the Federal Republic of Germany. The library is responsible for the collection, processing and bibliographic indexing of all German and German-language publications issued since 1913. This task is based on 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 until now. Only in the next few weeks (the expectation is that the new law will come into effect in the first half of 2006) will we get new legislation, which will enlarge our area of responsibility to include all types of net or electronic publications. And it is to be expected that in consequence of this new law, a number of regional libraries with legal deposit responsibility for specific regions will get a completely new legal foundation for their collections of digital publications.

Libraries will have to be prepared for this new situation. This is why in the last few years Die Deutsche Bibliothek has started a lot of initiatives to promote the long-term preservation of digital publications in Germany. A number of basic 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 be submitted once.

Based on 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' taskforce explored and established

the conditions necessary for Die Deutsche Bibliothek to become a deposit library for online publications as well.

Since 1998 online dissertations and theses (45,000 so far) have been collected, archived, and made available on a document server. Electronic periodicals have been collected since 2000, and since the year 2001 Die Deutsche Bibliothek has been operating a submission interface for online publications. During the submission procedure, DDB also asks for technical metadata relevant for preservation purposes. This has to be a compromise between the workload acceptable to publishers participating in voluntary submission, and the extensive requirements of the processes in the deposit system for future preservation. Other steps in the coming years are special developments for example for newsletters and for the retrospective ingest of different materials and collections. Furthermore, Die Deutsche Bibliothek has participated in the European Nedlib-project (Networked European Deposit Library) to adopt the OAIS-model (Open Archival Information System) and to develop workflow-suggestions for the integration of library procedures.

Other relevant experiences:

- DDB has built up additional experiences through 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).
- Another activity covers the development of a persistent identifier infrastructure for Germany. The use of persistent identifiers is the only possibility to guarantee that a digital object can be addressed permanently in the Internet. Embedded in the project "EPICUR - Enhancement of Persistent Identifier Services - Comprehensive Method for unequivocal Resource Identification" persistent identifiers become a part of a metadata framework for electronic publications [2]. DDB has chosen the Uniform Resource Name (URN), which is a Uniform Resource Identifier with the term ‘urn:’ preceding the rest of the name, and which serves as a permanent designator for a resource independent of location. URNs are persistent (i.e. they never change) regardless of whether or not the resource’s physical location changes. The purpose of URNs is to identify a single resource, and it alone, for the duration of its existence. But it should be noted that a resource can have a number of URNs allocated to it.

National initiatives

Germany has a federal structure with important elements of self-government in the states , especially concerning the education system and the science and research sector. The existence of a lot of regional libraries with legal deposit responsibility for their regions is another part of the federal structure. And considering the importance of the task of long-term preservation within the federal structure of Germany it is obvious that the approach to a successful solution to these issues in Germany must be cooperative. This primarily concerns the organizational aspect, but there are also a lot of practical and technical reasons why we are trying to distribute the responsibility for collecting the electronic objects in the Internet. With this background it is clear that we need partners in order to implement a long-term

preservation strategy and infrastructure in Germany in two directions: organizational and operational.

There are two initiatives – embedded in a lot of smaller projects and initiatives not all of which are mentioned here - through which Germany is trying to approach the problem of long-term preservation:

- From a more general and organizational perspective nestor (Network of Expertise in Long-Term Storage of Online Resources) was established with the goal of building up a platform of competence for sharing knowledge and experience in the field of long-term preservation and to exchange experts and expertise between different types of cultural heritage institutions.
- Die Deutsche Bibliothek, with partners, was given the task of building up a long-term archival system based on OAIS as a practical aspect of the general scope of kopal (Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen / Co-operative Development of a Long-Term Digital Information Archive).

nestor

Financed by the German Federal Ministry of Education and Research *nestor* [3], the alliance for Germany's digital memory, started in 2003 and will be completed in the middle of 2006. Under the leadership of Die Deutsche Bibliothek, there are several partners from the library area (Bavarian State Library, Göttingen State and University Library), media centers (Computer and Media Service of Humboldt University, Berlin), archives (Bavarian State Archives – Head Office) and museums (Institute for Museum Research, Berlin) on board. Additionally, on the advisory board there are publishers, representatives of science and technology, museums, archives, libraries and universities and also members of culture and politics and research institutions / computing centers.

The central aim of the nestor project is to bring together available knowledge, people and expertise on long-term storage of digital resources as a starting point for a future alliance for Germany's digital memory. Therefore the project:

- creates a network for information and communication about present and future long-term preservation (LTP) activities in Germany;
- establishes a cross-sectoral community to promote and support LTP activities and to raise awareness in society;
- triggers synergies between on-going activities in Germany and cooperates with international partners and projects;
- developes strategies for coordination of LTP activities in Germany;
- proposes a long-lasting organisational model to continue the service as a network of excellence after the end of project nestor in 2006.

In detail, the tasks and measures are:

- collecting and presenting information;
- consolidating areas of expertise and making them visible and available;
- promoting cooperation and supporting a common solution;
- preparing expertises on technical, organisational and legal issues;

- presenting models, putting them forward for discussion and encouraging widespread best practice;
- promoting standards and representing Germany on international standardisation committees;
- developing collection guidelines and selection procedures for the storage of digital sources;
- coordinating responsibilities for long-term tasks;
- raising awareness of the problem in specialist documentation circles and among the public;
- preparing a permanent organisation which coordinates and represents the concerns of long-term storage.

An important aspect – mentioned here as an example of the activities - are the working groups and the expert reports. At the moment the following groups are active:

- nestor Working Group on Trusted Repository Certification,
- nestor Working Group on Multimedia Archiving,
- nestor Working Group on Preservation Policies and Selection Criteria.

The expertises address the following topics:

1. electronic Journals;
2. perspectives of long-term preservation of multimedia objects;
3. development of a descriptive profile for a national long-term preservation strategy (Preservation Policy);
4. digital long-term preservation and the law;
5. study of the state of existing research data and raw data from scientific activities;
6. a comparison of existing archiving systems;
7. digitization and preservation of digitized material in German museums.

In the meantime most of them have been published and are available in the Internet [4].

kopal

kopal [5], which started in July 2004, faces the problem of long-term preservation from a practical perspective.Technically and organizationally a collaborative approach has been chosen. Financially supported by the German Federal Ministry of Education and Research, kopal is developing an innovative technical solution in the form of a reusable long-term archive for digital data. The solution is – after a market survey done in 2003/04 - based on DIAS (Digital Information and Archiving System), jointly devised by IBM and the National Library of The Netherlands in The Hague. However, for the purposes of the consortium, there was a need to define some special requirements, especially for the cooperatively and independently usable design of the software solution.

Important points in the requirements and the enhancement of DIAS are:

- remote access capability for partners at different locations;

- flexibility within the system to handle a wide range of formats and different metadata schemes of different (and in comparison to the Netherlands often smaller) partners;
- multi-client capability, handling of personal filestores (“lockers”);
- separation of DIAS and additional tools for handling ingest and dissemination in order to get a flexible and easily adoptable solution.

The system – like DIAS - is implemented in accordance with international standards for long-term archiving and metadata within the OAIS framework. The possibility of integrating the solution into existing library and information systems is a fundamental objective of the project, and is only possible through transparency by using open and dedicated interfaces.

Some other important complementary components in addition to the existing DIAS-system are:

- realisation of monitoring and steering functions to prepare for the long-term preservation of digital documents (as a starting point to incorporating preservation planning facilities);
- flexible data import and export functions based on the object description scheme METS (Metadata Encoding and Transmission Standard of the Library of Congress, USA), in an enhanced and specifically adopted form as Universal Object Format (UOF) (see figure 1).

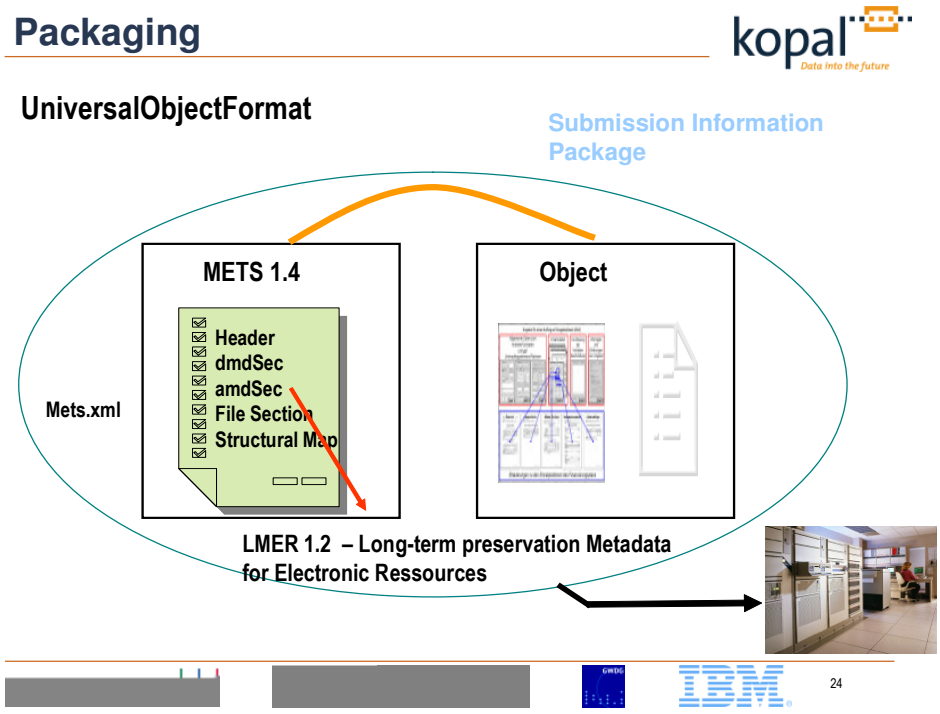

Figure 1: Universal object format

The requirements for a universal object format are described as a processing scheme in DIAS Core. In this arrangement, the object is processed as a Submission Information Package (SIP) and delivered as a Dissemination Information Package (DIP), using the OAIS model terminology. Within the DIAS

Core, a part of the data, called an Archival Information Package (AIP), is separated and put into a storage location (e.g. hard drive or magnetic tape). Special metadata are stored in a "data management" database, to which administrative access can be given.

For a fully functional strategy for the long-term storage of electronic documents, it is absolutely necessary to compile the appropriate technical metadata. Unfortunately, no standard for a suitable metadata scheme specifically for long-term archiving has been developed for a long time. Therefore, Die Deutsche Bibliothek has introduced its own scheme, called LMER (Langzeitarchivierungsmetadaten für elektronische Ressourcen / Long-term Preservation Metadata for Electronic Resources) [6], derived from a model at the national library of New Zealand. The mostly automatic extraction of technical metadata is based on results from the JHOVE-project (JSTOR / Harvard Object Validation Environment) and kopal is increasingly becoming an active counterpart and contributor in the software development (Neubauer and Wollschläger, 2006).

To get an open, and enhanceable solution, various types of partners are taking part in the project. And because of differing motivations, the partners decided to keep the aspect of system maintenance separate from development. The partner responsible for the operation of the system (the computing center Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen, GWDG) is hoping to gain experience with a well documented and scaled system, because it wants to attract further partners, who will use the benefits of the system for their own needs. On the one hand there is a neutral platform, where the developing partners Die Deutsche Bibliothek (DDB) and Goettingen State and University Library (SUB Göttingen) have to find common solutions for software and preservation procedures. On the other hand the system has to become capable of handling multiple users in a way which guarantees the independence of individual partners.

Within the project two of the project partners (DDB and SUB Göttingen) are putting digital material of all kinds into the long-term archive via batch processes. This ranges from digital documents in PDF, TIFF or TeX format to complex objects such as digital videos. After installing V2.0 of DIAS, a result of the first months of project-based software development including especially the realization of Universal Object Format capability, the project members are now starting the ingesting procedures in order to load multiple objects into the system. At the same time, some efforts have been made to establish a presentation system in DDB, which is based on a special caching area (here a server) that buffers used objects in an access area and delivers the objects rapidly according to user needs.

Regarding the software architecture, there is a separation between the core functionality of the archiving system and the environmental tools, which handle the homogenization and the transfer of digital objects into the system. This task was taken on by DDB and SUB Göttingen using a cooperative and modularized concept, based on JAVA-classes. The tools for building standardized Submission Information Packages (SIPs) and for importing them into the system come with an open source licensing (GPL) method. The free software "kopal Library for Retrieval and In-gest" (koLibRI), with which archival objects can be created according to the UOF, will be available for public testing and analysing from March 2006 [7]. For the presentation system the same principle applies: results must be independent from special dedicated solutions, based on well-defined interfaces, and open to other partners and systems (see figure 2).

The development partner for the enhancement of the DIAS V.1.0 to the DIAS-Core (DIAS V.2.0) is IBM Germany GmbH. This will ensure a professional adoption of software components and provide stable long-term support. The separation of a core functionality (DIAS-Core) demands well-defined and freely available interfaces for future partners.

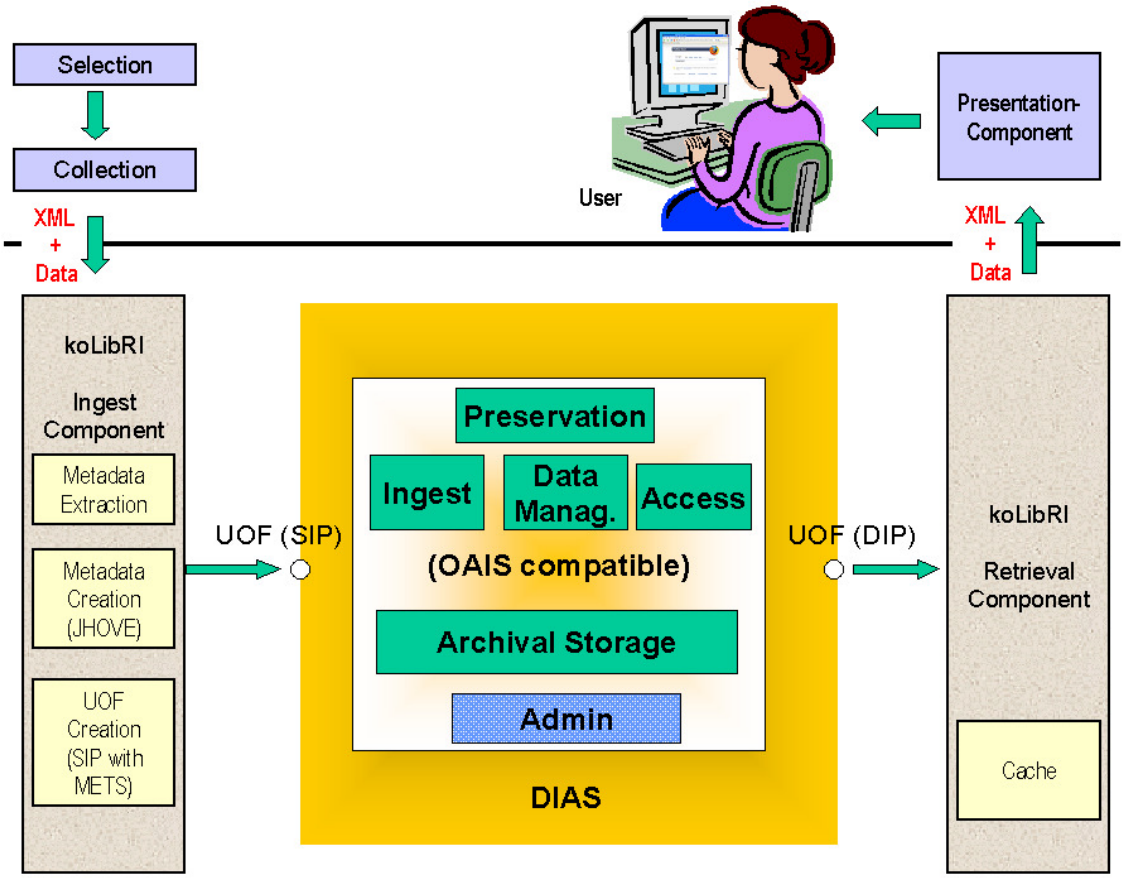

Figure 2: Architectural overview

Another important goal of the project is the development of business models in the sense of organized cooperation and dedicated licensing solutions, in order to deliver multiple and flexible solutions for heterogeneous partners. Therefore, kopal has integrated various partners at different locations from the outset. In the future, this long-term archive for digital information will therefore provide other institutions with the opportunity to keep their data available on a long-term basis. Consequently, kopal ensures the possibility of academic, business and administrative use extending beyond libraries. On the one hand, there is the possibility of a client having its own “locker” in order to use the system with a secure storage space under its own administrative control. This solution is especially appropriate for small organizations or ones with a small amount of material to be archived. On the other hand, there is the possibility of later use of the kopal solution by installing the DIAS Core, which can be run together with kopal tools developed and supported by kopal. Additionally, as part of the project a

working network / working group with the national library of the Netherlands was established in order to promote software innovations and strategic developments.

With the new architecture – sharing resources and spreading the use of the system – DIAS, in the form of the kopal solution, can become a central point in the worldwide search for solutions and strategies for preservation planning. The most important point in the project planning for the next 18 months is therefore the dedicated development of a detailed storage and service concept, the development of migration management tools and finally the finding, testing and systematic implementing of emulation strategies and tools.

Notes

1. www.ddb.de/
2. www.persistent-identifier.de/?lang=en
3. www.langzeitarchivierung.de/index.php?newlang=eng
4. www.langzeitarchivierung.de/modules.php?op=modload&name=PagEd&file=index&page_id=18
5. http://kopal.langzeitarchivierung.de
6. www.ddb.de/standards/lmer/lmer.htm
7. http://kopal.langzeitarchivierung.de/index_koLibRI.php.de

References

Neubauer, M. and Wollschläger, T. (2006), „Maschinelle Gewinnung technischer Metadaten für die Langzeitarchivierung elektronischer Publikationen“, *B.I.T. online*, Vol. 1/2006, pp. 37-40.