Date : 01/07/2008

Digital preservation:

Part 2: ***Preservation infrastructures***

Implementing a cooperative long-term preservation infrastructure solution for heterogeneous institutions – report on activities in progress in Germany

Reinhard Altenhöner
Deutsche Nationalbibliothek, Frankfurt/Main

Meeting:	**84. Preservation and Conservation, (PAC), Information Technology, IFLA-CDNL Alliance for Bibliographic Standards (ICABS) and Law Libraries**
Simultaneous Interpretation:	Not available

WORLD LIBRARY AND INFORMATION CONGRESS: 74TH IFLA GENERAL CONFERENCE AND COUNCIL

10-14 August 2008, Québec, Canada

http://www.ifla.org/IV/ifla74/index.htm

Abstract:

With regard to implementing the kopal solution (Co-operative Development of a Long-Term Digital Information Archive, please refer to http://kopal.langzeitarchivierung.de/index.php.en), the National Library of Germany and their partners serve and provide a solution for one of the remaining unsolved problems of our information society: Given the increasing number of electronic publications, we have a strong need for a reliable long-term solution in the form of a – from our point of view in terms of resource-sharing - cooperatively developed and operated archive for digital data. Now after some years the solution is available and is gaining operational capabilities – at the moment as a solution for the National Library of Germany and the Goettingen State and University Library – jointly combined in one system.

The initial goal of the kopal project was to develop a technological and organisational solution which offers a transparent integration into existing library systems, and in which reusability by memory institutions plays a critical role. So in the implementation of the system, international standards for long-term archiving and metadata were adopted. In this way, both sustainability and the ability to further develop the system are guaranteed.

But now the solution must become operational in a broader scope of responsibilities, partners and tools: So we've initiated the next step, which means, that we establish a consortium of different players with specific requirements for the long-term archiving of digital objects. In this team, we have a representative critical profile of heterogeneous partners, collections and needs.

The goal now is to integrate this different roles and functions in the long-term preservation architecture on a machine-based level in such a way that every user-driven institutional repository will be provided with versions of digital objects on demand. In the future, this will allow us to offer objects, which are accessible with a standardised client-viewer environment or with a presentation in a virtual environment, running on a remote platform. This means that we have to organize distributed secured ingest workflows, specific communication infrastructure on the level of objects and actors and last but not least we need a business model for all these types of services.

The presentation will give an overview of the development currently in process and tries to give an ambitious vision of some needs and musts for an international long-term preservation infrastructure.

Implementing a cooperative long-term preservation infrastructure solution for heterogeneous institutions – report on activities in progress in Germany

1. For some years now, we already have long-term preservation systems or archives in place, most of them – noted in statements on the own project - use the OAIS-framework as a scale for their own ambitions and in order to describe the basic architecture of the system. Another characteristic of these systems is the high degree of centralisation, which means that often we find solutions which were made for one single institution. For example DIAS[1], the Digital Archive Information System, developed by IBM for the National Library of the Netherlands[2] is an inhouse solution deeply integrated in the workflow organisation of the library and with no interfaces to external sources. Another precondition for those developments is that these achives are typically dedicated to a small number of deliverers of digital material. This guarantees homogeneous procedures, because with a small number of well known partners the need to normalise data and metadata information is not as great as it is with a huge number of partners, each having their own dedicated routines and workflow organisation. Most of these solutions are focussed on publications or static objects published on specific publisher sites in the WorldWideWeb.

Another example for this type of archives is Portico[3], which represents the model of a dark archive in the sense of a closed solution dedicated to fulltext articles from serials and journals. Every ingest into the archive is normalised into a well documented XML-based data-format before the digital objects are archived. The access to the preserved materials is restricted to those who own the respective right and follow the copyright rules of the publisher. Portico offers a last resort solution in the sense that titles only become available if a specific trigger event has occurred. In the meantime, portico has the complete reponsibilty to offer usable versions of the already ingested materials and make the right decisions

[1] The Koninklijke Bibliotheek, please refer to http://www-05.ibm.com/**nl**/**dias**/preservation.html

[2] Please refer to http://www.kb.nl/index-en.html

[3] Please refer to http://www.portico.org/about/approach.html in order to learn more about the basic policy of portico.

about concrete measures like migration activities. There is no transparency on events which have taken place inside the archive. So this strategy is comparable with an assurance policy which offers a guarantee for critical events.

In both cases we have backing organisations who take over full control of the whole process of ingest, of the preprocessing activities, of decisions on the policy for migration paths, on the control on dissemination and so on.

The systems work more or less independently from each other; small areas of cooperation have been determined, but are not yet operational. So the exchange of archived objects and work sharing is still an idea in order to make these activities more efficient and to save resources. For some deliverers the parallel approach (and their presence in both archival systems) is a dedicated goal in their strategy, because having two different systems is a better guarantee that the ingested objects will remain available over time.

In the past two years some progress was made to define and standardise metadata formats, but the practical use e.g. of the PREMIS-approach is still rare.

In a systematic perspective we can note the following observations:

- Operating long-term preservation systems are concentrated on big deliverers with automated data processing routines
- Most of the solutions are proprietary in a technical sense as well as being tailored to very specific user groups
- The number and relevance of well documented machine interfaces in this field is low
- The systematic decision and processing of digital preservation processes is hidden inside the “black box” of the archive solution: deliverers (producers like publishers or libraries as licensing partners) have no possibility to influence e.g. dedicated migration steps
- The need to normalise workflows and objects is high
- The interoperability between digital preservation achives is less

In this sense, the existing digital preservation infrastructure is dedicated to a small number of trusted long-term archiving repositories, which have the control and the responsibility on our digital heritage. The idea to establish a network of interacting sytems – the safe places network – was announced by the KB in 2006 and some steps have already been made towards an implementation.[4]

In Germany we were confronted with another challenge: Germany is a federated country which is characterised by a shared power on different regional/national levels of responsibility. Even in the area of libraries we have round about 35 libraries with a distinctive responsibility for the legal deposit in their respective region. Having this in mind, it becomes obvious that those libraries on the one hand have a task, which means that they have to identify, to collect, to index

[4] Oltmans, Erik and van Wijngaarden, Hilde (2006) The KB e-Depot digital archiving policy. Library Hi Tech 24(4):pp. 604-613, cited from the eprint version: http://eprints.rclis.org/archive/00009159/01/oltmans_vanwijngaarden_final_web.pdf

and archive digital objects, and on the other hand have difficulties to fulfill their task in a appropriate way because of a lack of adequate resources and methods.

Long-term preservation requires a large investment in personal and material costs. Therefore it's rather clear that digital preservation has to be done in a cooperative way which integrates the experiences of dedicated communities and their feedback. In a global perspective we have some approaches to come to cooperative shared activities e.g. for format registries or the automated extracting of meta-information.[5]

In 2004 we initiated our approach to develop a cooperative solution for long-term preservation on a national level.

2. The German project 'kopal' (Co-operative Development of a Long-term Digital Information Archive) had the mission to practically prove and implement a co-operatively built and used long-term preservation system for digital publications.[6] Within kopal, the partners have developed a technological solution for long-term archiving that includes not only the archiving and bit-stream preservation of digital documents, but also the implementation of preservation planning mechanisms (especially migration) for digital documents to ensure their accessibility in the future. Kopal is based on the DIAS solution of the IBM, originally developed for the KB. Kopal leverages the commercial system DIAS with an underlying commercial software set of IBM-standard software, which was extended especially for remote access, enhanced metadata administration and extended machine-readable interfaces. Additionally, in the project, an open source software JAVA library was implemented, used for automated ingesting routines (extracting of metadata, quality control, ingest and retrieval). Additionally object validation and metadata extraction software was integrated and amended.[7] The kopal Library for Retrieval and Ingest, koLibRI, is therefore important in the sense, that the reuse and the possibility to integrate the features in other system-environments has a crucial impact on the success of the complete solution. Furthermore, the software is used to migrate defined objects in the system in an automated workflow by governing the validation and access mechanism.

As of June 2006 the Deutsche Nationalbibliothek and the project partner, the State University Library Göttingen, is ingesting parts of their digital collections into the system. In mid 2007 the project was finalised and now the operative phase of the project has started. In a cooperation contract between the libraires, IBM and the data center GWDG (Gesellschaft für wissenschaftliche Datenverarbeitung), which drives the operational service of the archive, all partners have agreed to continue their work and to enable other institutions to join the consortium. The kopal archival system has been transferred into practical use and is about to be adopted by further partners from the library and heritage community.

[5] DFG-Rundgespräch "Workshop on Preservation networks and technologie", https://www.ibi.hu-berlin.de/.../tagungen/workshopjune07

[6] http://kopal.langzeitarchivierung.de/index.php.en

[7] We are referring here to the Harvard project JHOVE, please refer to http://hul.harvard.edu/jhove/ .

As stated before one major goal in the project was the development in a cooperative environment, which allows multiple institutions to participate on different levels of involvement. This means on the one hand the work sharing (data center being responsible for the operational service and the bitstream preservation, the libraries being responsible for all the aspects of preservation planning and the ongoing functional extension of the solution) and on the other hand the transparent integration into existing library systems and the reuseability through memory institutions plays a critical role. Considering the aspect of flexible reusability international standards for long-term archiving and metadata were adopted. In this way, both sustainability and the ability to further develop the system are guaranteed.

The digital preservation solution needs therefore to be embedded in the working environment and dedicated workflows, in which cultural heritage organisations collect, share, disseminate and present digital objects. Basically in kopal a distinction of system users between “clients” (in different stages) and those, who responsible for the complete system, was made. In this sense the possibilities to reuse kopal in a productive environment reach from the complete outsourcing up to the self driven inhouse system solution inside an institution. The last possibility is rather expensive in terms of funding and staffing. Therefore a more differentiated client-oriented solution was adopted:

Clients in the sense of account owners rent an account similar to the bank accounts we are familiar with. In kopal this account is a virtual area on top of the system that is independent from other participating institutions. This means that every participating institution has their own dedicated account, which can be administered for their own purposes. In consequence those organisations assume the responsibility to curate the digital content they collect e.g. from the Web. So the role is extended to the obligation, to run the ingest-service and especially the curation activities in their own responsibilty. An account-holding member uses the platform and additionally he is responsible for the normalisation and evaluation of data. Even the presentation of the archived objects is part of the task. Also, the planning, the conceptual preparation and the implementation requires additional steps such as the systematic migration e.g. of dangerous (or at least difficult) formats. This will take place together with other account-holders and needs investment in know-how, permanent monitoring and qualification of personal.

Clients in the sense of kopal-participants on the other hand assign the cited tasks to other institutions with the status of an account owner, who is responsible for the curation of digital objects and the services all around digital preservation. The participants are obliged to describe the policy which should be followed in the system for their own ingested entries, they select and describe the objects for the purpose of long-term preservation. From the perspective of a participant the solution makes sure that the amount of effort is reduced in comparison to the needs of a full archival system and at the same time it is possible to influence the rules and regulations in the archival system, to participate on the discussions and to take over dedicated responsibilities for specific tasks in the whole process of digital preservation in a cooperative working szenario.

These model of operation and organisation describes the range of possibilities and the potential options, where long-term preservation with this dedicated background could take place. The cost model basically developed within the kopal project is dependent on the degree of measures / services the leading organisation (the “account owner”) is willing to assign.

Advantages of this approach in sharing the tasks and the degree of responsibility are:

- resource sharing
- shared licensing costs
- optimised use of distributed knowledge

The kopal project has developed in its life time the basic functions to implement and fill in the roles and responsibilities described before. But especially on the area of automated communication between different systems and the practical level of operational organisation kopal still needs some addional development.

Technically this means that there is a need for an enhanced rights management in order to provide different levels of ingest and retrieval. On the other hand there is a need for a seamless integration into the system environment of the various partners. Another goal is to obtain more information about the costs and amounts of work for the introduction of long-term preservation processing into different types of organisations. And in the end, it will be possible to generate valuable estimates for a funding and investment model for a complete infrastructure solution.

The partners will offer a package of services, which allows reusers to choose between different levels of service and to customise the existing solution to their specific needs. Furthermore the partners deliver dedicated consulting and operational services. Identified positions / factors in the cost model are:

- Consulting & support
- Detailed planing
- Hardware extension, licenses
- Adoption / customisation of SW-components
- Ingest
- Operational service

After it became obvious that the technical solution in kopal is not detailed enough to address the different needs of potential partners - especially smaller institutions, which bring the demand for simple and integrated solutions – the planning for the next step starts.

3. The impact of digital preservation in the portfolio of libraries and comparable cultural heritage organisations is becoming increasingly important. If they fail in this area, customers will probably decide against them. In this sense they need solutions which allow them to care for digital objects in their area of responsibility. In the discussion with several organisations and institutions it has become clear that there is a need to have different models of implementation with a high degree of customisable options. And it was recognised that these institutions are motivated to become involved in the basic principles of digital preservation. The requirements for this engagement are very different and this means that the single services must be applicable to the user needs. By agreeing

on a consortium structure with a documentation center, different types of libraries and library service providers and a virtual consortium of other institutions basically located in the information infrastructure of science and research the range of potential requirements was strengthened in a way that it's rather sure that the project covers most of the needs of those organisations in a prototypic way.

Therefore the existing services should be enhanced to a real cooperative solution – not only in a technical sense, but in a operational / organisational sense. So the starting point to initialise the next step of development for the cooperative kopal solution is identified:

Based on the kopal results thus far, the partners wish to improve the practical reusability of the software development. In order to create a generic solution that can be implemented in many heterogenous environments and integrated as a part of the working policy of cultural heritage organisations, there is a need to develop an open concept with modularised service packages.

These are the general goals:

- Creation of a flexible long-term preservation infrastructure adapted to the needs of (smaller) cultural heritage organisations and their service providers
- Technical enhancement of the existing solution, conforming to the partners' requirements
- Implementation of a reusable process model and preparation of a handbook to introduce long-term preservation in (smaller) cultural heritage organisations

These are the key factors to establish the cited infrastructure:

- ➢ different models to realise ingest procedures

Starting point is the flexible software library koLibRI, which will be implemented as a full service solution at the partner institution, as an installation for remote use by some partners and as a service provided by DNB/SUB to some partners based on defined rules.

- ➢ interfaces

The kopal solution is an open solution in the sense that the project is strictly based on the idea of open interfaces, well documented by publication of the specifications in the web. On the other hand from the perspective of a software architect the solution is modularised and ready to be enhanced. Additionally, in the project it is necessary to extend the possibilities to address these interfaces and to govern them by external software applications. One task is e.g. to allow the transparent access from user systems under secured conditions and the automated exchange of objects including the complete set of information necessary to integrate these objects into the preservation policy of the archive.

In the end, we will have generic ingest and dissemination interfaces, which allow for an enlarged set of reuse scenarios.

- access and presentation

Ingested objects are under control of the preservation planning policy of the kopal solution. This means that regular migration processes will take place under defined circumstances with agreed measures and in a secure trusted environment. In case of a request, the kopal solution offer a version, which is accessible with a state of the art viewer. In order to avoid too much user traffic and to share resources by concentrating on dedicated tasks (here preservation activities) partner repositories will be provided with the “Use”-version in single cases, with parts of collections or with the complete content of the server in order to rebuild a service. BTW: the compliance with copy right regulations is one of the general requirements with high relevance for the consortium

Therefore we need a communication syntax / interface between the service layer of the partners and the kopal solution.

- Technical enhancement of the existing solution following the partner requirements

The existing solution should be extended in the sense that we need validation and metadata extraction tools for new formats like video clips or other multimedia objects. Therefore we will extend our JHOVE library. And we wish to integrate the aspect of normalisation before the ingest process takes place – this can be helpful to facilitate ingesting procedures and will reduce the amount of work to be done later, when those specific formats become obsolete.

Other issues are the implementation of complex migration scenarios and specific selection needs.

- Implementation of a reusable process model and handbook to introduce long-term preservation into cultural heritage organisations

In many organisations we find discussions revolving around the question of how the long-term preservation of digital material can be initiated. We think that there is a demand for a guideline in the form of a checklist which articulates the basic questions and helps to document the starting position in an organisation. Most of the organisations are specialised on selecting, indexing and presenting relevant sources and objects and it seems to be really difficult to introduce an integrated approach for digital preservation. The following subtasks have been identified:

  - Analysis and documentation

In the scope of the project there is the analysis of the user requirements and a need to filter the general recommendations and rules. We need a consolidated and summarised catalogue of key questions which help to identify the basic needs. Special attention should be given to the sustainable integration of long-term preservation relevant processes and workflow implementation.

- Business model

A key factor in creating successful models of cooperation is the existence of an appropriate business model. In this sense there is an explicit need to define costs and models, to identify the individual cost factors and to define the service levels which can be offered.

- Service

The definition of a service offer is one other important challenge within the project in order to guarantee the continuity of the service and to establish this service scenario for cultural heritage organisations.

- Recommendations

There are frequent requests for a catalog of practical advice and a recommended course of action in the form of a handbook.

4. The topic long-term preservation has raised a great deal of attention worldwide, but the attention is out of proportion to the quantity of real operational solutions. Probably one of the reasons is that the amount of effort to set up those systems is high and remains high, because continuing activities are necessary to keep digital objects available for a long time. A common characteristic of those solutions is therefore the degree of complexity and the sizeable investment.

On the other hand the expectations of users / customers grow and they ask for policy, regulations and quality guarantees for certified services. In this sense, the memory institutions (including the science & research infrastructure) wish to influence the policy of long-term preservation archives – they collect and select relevant material and they become involved as a part of the long-term preservation infrastructure. In parallel the range of objects and the requirements in terms of durance, quality and availability of research data grow too and the institutional pilars in the information infrastructure like archives, data centers or even libraries have to guarantee the sustainable availability not only of digital information objects but the permanent existence of linking information between data and corresponding publication objects, between comprehensive entities like authors. And those elements of a sustainable, permanent available infrastructure for information depend from the level of cooperation and resource sharing we will reach.

In this respect the complete outsourcing of the long-term preservation solution to dedicated service-providers reduces the amount of effort for individual institutions, and on the other hand, they become involved by finding possibilities to measure the quality of these services, defining their own criteria and their methods and measures to control the services.

In a shared, cooperative digital preservation environment we need therefore customised services ready to become integrated into the existing workflow procedures. And we need well established exchange mechanisms for objects and metadata. And we need standards and a certification infrastructure to certify services and service-providers in a sustainable and trusted process. Both directions, the ingest process and the dissemination process have to be addressed in a machine-usable way. The providers of long-term preservation services benefit from the input of memory institutions, because they assume the task to select and index the relevant digital objects.

The landscape of long-term preservation will underscore the role and importance of cultural heritage organisations. They provide sustainable information on objects and ensure permanent access to objects. In a cooperative environment with dedicated tasks for the different players in the arena, the access has to be organised from the day by day access, to actual materials, up to the request for old historic formats in the original technical application environment. And those requirements can only be served in a cooperative, well defined, shared environment with distributed responsibilities. Our effort will be a contribution for this global digital preservation framework.