BABS2: A NEW PHASE, A NEW PERSPECTIVE IN DIGITAL LONG-TERM PRESERVATION – AN EXPERIENCE REPORT FROM THE BAVARIAN STATE LIBRARY

First author
Tobias Beinert

Second author
Dr. Markus Brantl
Bavarian State Library
Ludwigstraße 16
D-80539 München
{beinert, brantl, kugler}
@bsb-muenchen.de

Third author
Anna Kugler

ABSTRACT

BABS is an acronym for Library Archiving and Access System (Bibliothekarisches Archivierungs- und Bereitstellungssystem), which constitutes the infrastructure for digital long-term preservation at the Bavarian State Library (BSB). During the two-year project BABS2 funded by German Research Association (DFG) BSB focuses together with the Leibniz-Supercomputing Centre (LRZ) on advancing its organizational and technical processes under the aspect of trustworthiness according to the nestor criteria catalogue. Important achievements are e.g. framing an institutional policy for digital preservation including local, regional, national tasks of a large-scale research and archive library, conducting and evaluating a survey concerning the archiving requirements of all BSB departments, documenting the ongoing archiving processes, introducing an appropriate quality management and improving the scalability of the preservation system. Additionally BSB participate in different national and international committees.

This experience report sheds light on the various organizational and technical aspects which have to be taken into consideration when enhancing an existing infrastructure for digital long-term preservation.

1. INTRODUCTION

Today the Bavarian State Library (BSB)[1] as universal and international research library manages the largest digital archive for cultural heritage in Germany, now containing more than 380 million files with a total amount of 218 terabyte (June 2010). Several mass digitization projects, including the public private partnership with Google, as well as different web archiving and audio digitization projects contribute to the considerable scale and variety of resources in the archive. Not only these huge dimensions, but also new political responsibilities (e.g. the inclusion of online publications of public authorities in the legal deposit law) require several organizational and technical enhancements of the existing archiving infrastructure.

Responsible for long-term preservation within BSB is the Munich Digitization Center/ Digital Library (MDZ)[2], which during the last years has set up a "Library Archiving and Access System" (Bibliothekarisches Archivierungs- und Bereitstellungssystem, BABS[3]) in collaboration with its strategic partner, the Leibniz Supercomputing Centre[4]. Current preservation responsibilities of BSB include:

- Digitized books produced by the BSB or by commercial partners
- Born-digital documents delivered according to the current legal deposit act (e. g. governmental publications of the Bavarian state and of other German governmental institutions)
- Digital resources (e. g. websites, open access publications) belonging to the virtual library of BSB's special collection fields

as well as all other electronic media produced or licensed for use by the BSB.

For digitized materials the MDZ has developed, implemented and optimized a production line for all the relevant processes, e.g. preparation, scanning, metadata enrichment, delivery, and archival storage among others. A self-developed software-tool (so-called ZEND) fosters this workflow and makes it possible to cope with mass digitization and preservation of a wide range of materials from medieval manuscripts and incunabula, to journals and newspapers, as well as photographs and audio documents [1].

2. THE PROJECT BABS2

The current project BABS2, funded by the German Research Association (DFG)[5], faces the challenge to consolidate and improve the existing architecture for digital long-term preservation, integrate it into the overall organization of the library and adjust it to newly upcoming requirements. The aim is to build a trustworthy and scalable digital archive as part of a national network for digital preservation.

[1] www.bsb-muenchen.de

[2] www.digital-collections.de
[3] www.babs-muenchen.de
[4] www.lrz-muenchen.de
[5] German Research Association (Deutsche Forschungsgemeinschaft): www.dfg.de

The ongoing improvement processes cover organizational aspects such as designing a digital preservation policy for the BSB, developing new workflows, documenting the existing workflows, as well as technical aspects such as re-structuring the present storage system and introducing (periodic) virus- and checksum scans.

During the project, experiences with innovative methods (e.g. preservation planning, self-auditing based on criteria for trustworthiness) will be made in the fields of organization, evaluation and improvement of digital preservation. In accordance with nestor[1] and in collaboration with the German National Library (Deutsche Nationalbibliothek, DNB)[2] and regional libraries, models for national cooperation will be developed.

3. ORGANIZATIONAL ENHANCEMENTS

3.1. Consolidation of the digital archive

The digital archive of the BSB was established out of the need to store the rapidly growing amount of data beginning with the first digitization projects in 1997. Now, more than ten years later, further consolidation of the organizational and technical infrastructure inside the overall institutional framework of the library is necessary.

A long-term-preservation unit inside the Munich Digitization Center / Digital Library was established already in 1999. With regard to the changing organizational structure of the BSB its tasks and responsibilities were further clarified. At present it is responsible for the connection to the digital production as well as for research and development.

A first milestone of the BABS2 project was the design of a policy which clearly defines the aims of the digital archive of the Bavarian State Library as one of the most important cultural heritage institutions in Germany, as the archive library of Bavaria, and as the head of the Bavarian Library Network. Besides a concise mission statement, it states the reasons for BSB's responsibility in the field of long-term preservation and tries to shape a basic profile for its collection and archiving duties in the digital world. Furthermore it comprises an explanation of the general principles which BSB adheres to (e.g. provision of customer-oriented digital services; ensuring trustworthiness; long-term preservation as a cooperative business etc.). The existence of such a policy is itself also one main criterion of trustworthiness according to the nestor criteria catalogue [2]. A first draft has been completed, it is now up for discussion by the responsible departments of the library and has yet to be adopted officially by BSB's head office.

The preparation of a written mutual agreement between BSB and LRZ in digital long term preservation formed a next important task for organizational consolidation in the BABS project. Since 2004 both institutions have been working together in several projects. Building up a jointly operated technical infrastructure and transferring into routine business presented a central milestone in digital long term preservation for both sides. In the context of the BABS2 project the LRZ performes important scalability tests (see 4.3) which will be the basis for a refinement of the Service Level Agreements (SLAs) between BSB and LRZ.

A further consolidation step of BSB's digital archive was the examination of the existing archiving workflows, including those for the ingest of monographs and periodicals, the workflow for digital materials, as well as process models. In addition to a detailed documentation of all current processes and activities, which provide the basis for trustworthiness, we developed new workflows e.g. for legal deposit and web archiving.

To prepare for future tasks in the library and the Bavarian Library Network we are conducting an inventory survey/stakeholder analysis for all major departments of the library and the partners of the network. Our aim is to review the digital material available and to detect present and future requirements for digital long-term preservation. The results will lead to the design of adequate organizational and business models for long-term preservation at the BSB and the Bavarian Library Network.

As part of our quality management we organized a workshop together with our project partners from LRZ, in order to self-evaluate our preservation architecture. In a first step we reviewed our digital archive according to the concept of trustworthiness set out in the nestor-criteria catalogue [2] using the following assessment scale:

- 1 = conception
- 2 = in process of implementation
- 3 = completely fulfilled

Many of the criteria were well fulfilled, such as all actions are based on legal and contractual regulations (e.g. controlled access to the digital documents) or the definition of necessary metadata, but other criteria concerning the organizational structure of the digital archive (e.g. distributed responsibilities over different departments and institutions) need further enhancement.

In a second step we applied the self-audit method of DRAMBORA (Digital Repository Audit Method Based on Risk Management)[3] to assess our digital archive. As we especially wanted to review the transfer of our digitized images to the storage at LRZ we focussed on a risk assessment of integrity and authenticity in the areas of ingest and storage. The risks we could identify in this area are in great parts already covered quite well by

[1] www.langzeitarchivierung.de
[2] www.d-nb.de
[3] www.repositoryaudit.eu

LRZ's own institutional risk management programme, but now have to be specified in greater depth for our joint preservation activities.

3.2. Cooperation activities

The long-term-preservation unit of BSB takes part in national and international collaborations, e.g. in committees such as the German competence network for digital preservation nestor and standardization working groups at DIN/ISO.

Within the framework of nestor[1] and in collaboration with the DNB and regional libraries, BSB contributes its share to the very challenging task of developing cooperative models for long-term preservation in the federal state of Germany. The activities in this area include amongst other things the planning of exchanging information packages in both directions between the BABS-system and DNB's system kopal as well as the participation in the LuKII -project[2] which aims at setting up a LOCKSS-network for Germany and testing the interoperability of that network with repositories and archival systems. With the Library of Congress BSB established a private LOCKSS network in order to test the exchange of electronic official publications.

Furthermore BSB is actively involved in the working group for a National Hosting Strategy within the framework of the Priority Initiative "Digital Information"[3] by the Alliance of German Science Organisations. A first result of these activities was the publication of a final report *Ensuring Perpetual Access* by Charles Beagrie Limited on the establishment of a federated strategy on enduring access and hosting of digital resources for Germany in March 2010 [3].

4. TECHNICAL ENHANCEMENTS

4.1. Enhanced AIPs, integrity and authenticity

For digitized books up to now bibliographical and structural metadata is saved with the digital objects. Due to new requirements regarding the heterogeneity of digital resources of a universal library (manuscripts, rare books, special collections) and the growing complexity in the technical production we decided to store further technical information for the singular image. We have developed an additional workflow to generate and save further preservation metadata (technical as well as event metadata) on the image level according to the PREMIS standard.

The technical metadata is extracted with jhove, but we are also evaluating the possibility of using FITS[4], because this tool integrates different extraction and validation services which comply better to our requirements. FITS is in comparison to jhove e.g. able to extract information about the ICC colour profile, which we need for further possible migration actions. Event metadata is saved in the process of ingest (e.g. creation, validation results, enrichment, normalization/migration actions). The technical metadata as well as the event metadata is stored in xml-files next to the digital object.

In a parallel effort the structural metadata of our digital objects is revised, so we will be able to tie an enhanced and enriched AIP which includes explicit metadata for long-term preservation.

Our aim is to perform virus checks and scans of the generated checksums at certain points of the established archiving workflow and store this information inside the AIP. We tested the required time and processor performance of these checks, which showed that generating as well as scanning a certain amount of checksums can be included into our digitization workflow without violating the on-going processes. We identified the critical stages of the archiving workflow and fixed certain points of time where the checksums can be generated and accordingly checksum and virus scans can be performed. All information is stored in the newly designed AIP of our digitized books. In terms of risk management we develop concepts on how to proceed in case of an integrity violation (see 3.1.).

4.2. Preservation planning

In a joint workshop with the TU Vienna in 2009 we have designed a preservation plan for format migration of a selected collection of digitized books (16th Century Printings) using the PLATO tool which supports the process of decision-finding and documentation [4]. We considered the option of migrating the selected digitized images from TIFF to JPEG 2000 as new archive format. Following the planning workflow of PLATO we first of all defined our requirements according to the digitization standards and preservation policy of the BSB. Some of our main requirements were e.g. to keep the resolution and the ICC colour profile of the image after the migration, to reduce storage costs or to allow the creation of full-text (OCR).

In a 2nd step we tested different alternatives of migrating from TIFF to JPEG 2000 with several open source tools, which showed different outputs. Concluding we evaluated the results and built the preservation plan for our collection: the alternative of "keep status quo" excelled over the other possibilities and was thus our recommended preservation action. According to our changing requirements and the developments of new or improved tools we need to review our preservation plan.

During this workshop we gained the necessary methodical skills to pursue further preservation planning for other collections, as for example for the legal deposit which will focus on pdf to pdf/a migration.

[1] www.langzeitarchivierung.de

[2] http://www.ibi.hu-berlin.de/forschung/digibib/forschung/projekte/LuKII

[3] http://www.allianzinitiative.de/en/core_activities/national_hosting_strategy/working_group/

[4] http://code.google.com/p/fits/

4.3. Scalability

In the BABS 2 project we cooperate with the Leibniz Supercomputing Centre to test the scalability of our digital archive. Tests with different storage management systems (SAM/QFS, TSM/HSM) showed that the huge amount of digital data BSB produces can't be handled easily by the well-established software. Together with previous scalability experiences (e.g. migration of storage media, handling the data of the Google-project [5]) it became obvious that a complete re-structuring of the storage system by virtual units should be done in order to allow further growth, improve access and performance, as well as allow migration of storage media which does not interfere with the daily routine.

4.4. Perspective: Introducing a new technical solution for preservation

Due to risen requirements, large scale of data, diversity of resources and a broader archiving focus, it is necessary to introduce a new, more robust system with long-term preservation functionalities. The above described experience and the gained knowledge prepared us for the introduction of the new archiving system "Rosetta".

The Digital Preservation System "Rosetta" was developed by ExLibris in a partnership with the National Library of New Zealand. Rosetta will enhance the technical infrastructure and its associated preservation workflows at BSB. The existing workflows will be improved, unified and consolidated into one single system. By introducing Rosetta, the BSB is on the way to apply several new features regarding long-term preservation. It offers e.g. a detailed risk analysis for each file, which provides the basis for the new preservation planning module. Preservation actions can be performed on a selected set of files according to the beforehand defined preservation plan.

The open platform architecture of the new system allows an easy interconnection with different external systems via customized API/SDK developments. All these efforts and activities improve the OAIS compliancy, the trustworthiness, scalability and robustness of BSB's long-term preservation activities substantially.

The introduction of Rosetta takes place in a pilot phase. During this time the specifications for the transition of three designated workflows into the system are designed, implemented and tested:

- Digitized objects
- Legal deposit
- Webarchives

According to these workflows several external systems and tools need to be integrated and tested. At the end of this phase Rosetta should be switched over to routine business and gradually the other workflows should be adapted to Rosetta. The already archived objects in the existing architecture will be migrated to the new archival system step by step.

In a second stage after introduction, libraries out of the Bavarian Library Network will join the BSB in using the system.

5. REFERENCES

[1] Brantl M., Schoger A. (2008): Das Münchener Digitalisierungszentrum zwischen Produktion und Innovation, in: Information - Innovation – Inspiration The Bavarian State Library. 450th Anniversary, p. 253–280.

[2] nestor working group for trusted repositories and certification (2009): nestor criteria – catalogue of criteria for trusted digital repositories, Version 2 http://nbn-resolving.de/urn:nbn:de:0008-2010030806

[3] Charles Beagrie Limited in association with Globale Informationstechnik GmbH: Ensuring Perpetual Access: Establishing a Federated Strategy on Perpetual Access and Hosting of Electronic Resources for Germany http://www.allianzinitiative.de/fileadmin/hosting_studie_e.pdf

[4] Kugler, A., Brantl M., Beinert, T., Schoger A., Kulovits H., Rauber, A. (2009): From TIFF to JPEG 2000? – Preservation Planning at the Bavarian State Library Using a Collection of Digitized 16th Century Printings, in: D-Lib Magazine, Vol. 15, Nov./Dec. 2009 http://www.dlib.org/dlib/november09/authors/11authors.html

[5]Wolf-Klostermann, T. (2008): How to cope with 300.000 scans a day. Managing large scale digital collections in practice – the Bavarian State Library and the Leibniz Supercomputing Centre approach the next level of mass digitisation, in: Archiving 2008, p. 272 – 274.