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.

Dr. Thomas Wolf-Klostermann and Dr. Bernd Reiner

Rapid increase in a short time

Since the foundation of the Munich Digitisation Centre in 1997 the Bavarian State Library became one of the major content providers among libraries in German-speaking Countries, now hosting more than 25,000 volumes and approximately 10 Million pages. Since 2004 archival storage is being done using the powerful infrastructure o the Leibniz Supercomputing Centre. The large digital collection will increase rapidly within the next years since several important new digitisation activities have been started during the year of 2007. First of all there is to mention the decision to adopt automated scanning technology and to use it even for the older holdings, e. g. books of the 16th to 18th century. Aim of the so-called VD16 project is to digitise the complete holdings of German books printed in the 16th century. This activity is funded by the German Research Foundation. Till 2009 more than 7.5 million pages of 16th century printings will be digitised and published online. Currently three automated scanning devices are in use for sixteen hours a day. The second important project is the public-private partnership of the library with Google. More than one million books out of the copyright-free holdings of the library are going to be scanned by Google within the next years. They will be hosted by Google Books as well as inside the Digital Collections of the Bavarian State Library for free access. Apart from these two large scale digitisation projects a lot of small activities contribute every day to a rapid increase of the library's digital collections. One example is the very successful "*Digitisation-on-Demand*" service, which offers the user the opportunity to order a digital copy of almost every single work out of the library's depository. Another examples are the growing electronic collections of medieval Latin and German manuscripts, of historical maps and incunabula. Putting it all together a daily amount of 300,000 scans has to be processed, ingested into the library's repository, published on the Web and stored in an adequate archival system. It requires several different activities and efforts to cope with this large amount of data. Together with the Leibniz Supercomputing Centre, the Bavarian State Library had set up a suitable technical infrastructure.

Main tasks

The infrastructure now in use was designed to fulfil four main tasks: 1) to manage digitisation jobs, 2) to do all necessary image processing, 3) to provide WWW presentation formats including bibliographic metadata and structural information, 4) to ensure the archival storage and the long-term preservation of all the data. Such heterogeneous tasks can only be realised with a heterogeneous system architecture comprising several different elements of hard- and software. The basic technical platform of the Munich Digitisation Centre is the so-called ZEND ("*Zentrale Erfassungs- und Nachweisdatenbank*") repository system; based on Open Source Software (e. g. Linux Operating Systems, MySQL Databases, PHP and PERL based Web interfaces and free image processing tools). Almost all the features consist of command-line based programmes and of comfortable, tailor-made web-interfaces. The basic features of the ZEND corresponding to the four main tasks listed above have been in productive state in the Munich Digitisation Centre of the Bavarian State Library since spring 2004. By the end of 2007 more than 24 M files (45 TB data) had been processed successfully by this system and stored inside the Storage System of the Leibniz Supercomputing Centre.

1) Job Management – efficiency through automised tasks

It doesn't make much difference weather scanning is done inside the library's own digitisation centre or by external commercial partners (e.g. Google) – the effort for the administration and logistics is almost the same. The book has to be identified, picked out of the depository, transported to the scanning facilities of the library or those of the commercial partner, digitised, returned (physically and electronically) and post-processed. The more standardisation and automatic processing is involved, the fewer mistakes could occur during this process. Most of the tasks are being done time scheduled and without human interaction. In the beginning the automatic import of the bibliographic record from the local catalogue system via Z39.50 takes place. A process slip with barcode is being created to accompany the book during the whole process. It contains the basic scanning parameters, e. g. the scanning device to be used, the desired resolution, the colour settings. The process slip allows the documentation of every single step inside a workflow database. After the automatic allocation of a unique identifier and of a URN the manual or automatic scanning process starts, followed by automatic image processing, web publication and automatic archiving of all the images. If the book was ordered by a customer through "Digitisation on demand" services, delivery and billing follows.

2) Image Processing and WWW publication

After downloading the digital master images from the temporary deposit of the scanning facilities, image processing starts. The basic output format is JPG. A PDF version could be created optionally. Every scan is generated in two or three sizes for different zoom levels (at 100%, 150% and 200%,

depending on the initial size of the book). An XML index file is been written for presentation purposes. Depending on scope it comprises either a TEI- or a METS-Structure. The JPG files and the XML index file are the basis of a standardized web output. Depending on the presentation scope and the context of the digitisation, an additional manual indexing takes place. For this, a web-interface based TOC editor was designed. It allows the easy enrichment of the XML index file with relevant structural information such as identification of the title page, chapter titles, images and foldouts. An OCR processing will be part of the standard workflow soon. Since not every work is suitable (e. g. older maps, manuscripts, handwritings, older German printings), an automatic pre-selection by media type and age has to be implemented.

The majority of the books digitised by the Bavarian State Library is freely accessible on the Web. In some cases copyright restrictions prevent from free access, In this case access can easily be limited to in-house use at the public internet PCs of the library's reading rooms.

3) Access – Multiple ways lead to electronic resources

When post-processing is finished, the URN and the URL to the digitised book are being reported back to the local catalogue system. This provides the opportunity to approach the digitised content with just a few clicks. At the same time the bibliographic information of the book is being added to the digital collection's search index. The OCR generated full-text of the book (if existent) is being allocated for full-text search service. In the end the user can choose between different ways of accessing the digital content. For instance he can use search engines on the WWW (e.g. Google or the WorldCat). Alternately he can search the library's local catalogue system, where a link inside the bibliographic record will lead him directly to the digital object. In addition the URN reference system of the German National Library is used to provide a persistent identifier for each object. The third option is to browse or search inside the Digital Collection's Homepage (http://www.digital-collections.de)

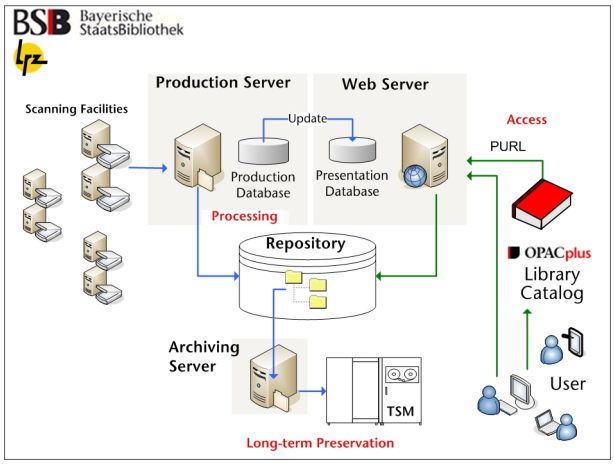
4) Archival Storage and Long-term Preservation.

In the background a TSM/HSM storage system based on a tape library is used to ensure archival storage. The incoming repository content is stored automatically in this archival system the very same day. Each digitised volume is been kept as an uncompressed master copy together with the complete bibliographical metadata and basic technical metadata information. Put together this makes a "self-explaining" archival information package which remains usable even in case of loss of all external reference systems (e.g. the database or the local catalogue system).

Hierarchical Storage Management (HSM) is used to economically store large amounts of data

by integrating several storage systems with different quality of services into a single file system view. Thereby the files are automatically and transparently migrated between the storage layers according to some definable rules. Because HSM file systems typically use magnetic tapes as final storage layer, the amount of data which can be stored is virtually unlimited. However, especially when HSM is used for long-term archiving with a rapidly and constantly increasing number of files, the performance of meta-data operations (e.g. identifying files for migration) could become a critical issue. This leads into several additional efforts necessary to keep the whole system manageable. Due to a very efficient architecture of the archival system every single stored file can be located and retrieved within an average time of 2 minutes.

Because only widespread and well documented file formats (TIFF, JPG, PDF/A and plain text files, e. g. XML) are being stored in the archive, other preservation activities (e. g. format migration and emulation technologies) are not yet in the focus of the activities, but could easily be implemented if needed. The first hardware migration of the complete data stock of the library was completed successfully by January 2007 (then 42 TB).



Large scale collections – large scale challenges

The basic characteristic of the described architecture is its volume oriented approach with a separation of bibliographical and structural data. While the basic object information and bibliographical metadata are stored in a database system, the structural information and a copy of the basic biographical information is kept in a XML document inside the repository directories. Browsing access involves just the directory. Only search operations connect the database. This separation allows increased performance even with a large number of objects and a very intensive public use.

The greatest challenge was to implement the existing services inside an infrastructure capable to handle more than a hundred times as much data. Since the beginnings of mass digitisation the technical infrastructure was split up between library (digitisation facilities, web servers) and computing centre (archival storage), bound together by a fast gigabit internet connection inside the *Munich Scientific Network* to ensure the daily data transfer via TCP/IP. Because of the huge amount of scans to be processed through the big digitisation

projects (VD16, cooperation with Google) the decision was made to re-install the whole server and archive infrastructure inside the computing centre – next to the archival storage systems and connected to the library by a fast internet wire. Because of the large amount of data the most important thing is to provide enough disk space as well as sufficient CPU capacity for the handling of all operations. For this a Network Attached Storage (NAS) System with a storage capacity of 9 Terabyte (primary storage) was installed. A cluster of 35 servers, each with a Quad Core CPU with 2.66 GHz, 80GB RAM and an expandable disk system was configured to ensure image processing as well as permanent provision of the WWW presentation copy. If needed, further devices can be added at any time to improve performance. A MySQL database serves as the central reference system.

Of absolute importance is the capacity of the existing infrastructure to process the incoming data the very same day. Every delay would be hard to make up since digitisation takes place on 5 days a week. Under this circumstances library and computing centre are well prepared to cope even with 300,000 digitised pages a day – the maximum assumed input when from all actual digitisation projects.

.

Institutional Information

*The **Bavarian State Library** in Munich is one of the leading research libraries comprising one of the largest collections of rare manuscripts, old printings, incunabula and maps in the world. Till today, more than 25,800 volumes have been digitised, published online by the library's Digitisation Centre*
*The **Leibniz Supercomputing Centre** sustains a powerful communications infrastructure called the Munich Scientific Network (Münchner Wissenschaftsnetz, MWN) and a competence centre for data communication networks, including storage and archiving systems of high capacity. It supports the Bavarian State Library in its long-term preservation activities since 2004.*
The Bavarian State Library: http://www.bsb-muenchen.de and http://ww.digital-collections.de
Leibniz Supercomputing Centre: http://www.lrz-muenchen.de