11 (2008) 3
AUSGABE

Reprint

D 52614
ISSN 1435-7607

B.I.T.online

Bibliothek Information Technologie

Zeitschrift für Bibliothek, Information und Technologie
mit aktueller Internet-Präsenz: www.b-i-t-online.de

Mass digitisation workflow management

Book and digital scan workflows

By Martin Baumgartner, Michael Beer, Berthold Gillitzer, Rosmarie Leichtl, Gabriele Meßmer, Karsten Trzcionka, Thomas Wolf-Klostermann und Wilhelm Hilpert

Mass digitisation workflow management

Book and digital scan workflows

By Martin Baumgartner, Michael Beer, Ingrid Bochmann, Berthold Gillitzer, Rosmarie Leichtl, Gabriele Messmer, Karsten Trzcionka, Thomas Wolf-Klostermann and Wilhelm Hilpert

■ In 2007, the Bavarian State Library concluded a much-acclaimed contract with Google relating to the digitisation of around one million out-of-copyright old books which are part of the Bavarian State Library's collection. The project will run for several years. One of the contract's stipulations is that the books are to be scanned within the Free State of Bavaria. It is based on the Bavarian State Library's pan-European „Call for Expression of Interest in Participation in the Negotiation Proceedings" published in Tenders Electronic Daily, generated widespread positive press coverage and triggered a lively debate among the librarian public.1

Abb. 1: Der Altbestand der Bayerischen Staatsbibliothek mag mit über einer Million Titel noch manchen unentdeckten Schatz für die Wissenschaft enthalten.

Public private partnership agreements are always formulated to satisfy each party's confidentiality requirements and, naturally, confidentiality has to be observed. This gives private sector enterprises the assurance that their know-how is protected, particularly the know-how relating to company processes and technology, which is often a key factor of competition. „Confidential" is the word that most aptly describes Google's relationships with its partner libraries. That is the reason why this article does not disclose any information about Google's scanning technology, quality assurance or internal processes.

This article views Google as a popular service provider which does, however, differ in two important respects from all other scanning service providers in Germany to date. Instead of receiving payment for its services, Google gets (or, to be more precise, retains) a copy of each digital scan and instead of a few hundred pages, it scans and processes many hundreds of books every day. Basically, the workflows described are typical of a mass digitisation work environment and they exist - in appropriately modified form - in the Bavarian State Library's other large-scale projects, such as the project to digitise 37,000 German 16th century prints.

From another perspective, the Bavarian State Library doesn't incur any direct costs for digitisation and it receives an identical copy, the "library digital copy" of the digital scan created by Google.

The preparatory phase

The reason for this article is the conclusion of the preparatory phase and the commencement of the productive phase of this mass digitisation project. Some people may consider a preparatory phase of more than 1 year to be too long, but it is actually quite normal in a project of this size and the project team members view it as short rather than long. First of all, Google had to provide the necessary human and material resources for a project of this size. For its part, the Bavarian State Library had to create the tools for the management, monitoring and reliable archiving of the books and digital scans.

The logistical challenge

Some people have no inkling of how big a challenge it is to make over one million books available and process them in a project with a relatively compact timeframe. In Germany, people often used to

1 Cf.: Klaus Ceynowa: Massendigitalisierung für die Wissenschaft – Zur Digitalisierungsstrategie der Bayerischen Staatsbibliothek. In: Information – Innovation – Inspiration. 450 Jahre Bayerische Staatsbibliothek. Ed. Rolf Griebel and Klaus Ceynowa. Munich: Saur 2008. p. 241-252.

talk about mass digitisation when a thousand books had to be scanned. Just for clarity's sake, this quantity was processed in less than one week in the Google project. The project marks the beginning of a new era of digitisation in German libraries - the era of industrial mass digitisation.
Although we said at the outset that we weren't going to interfere in „scanning service provider" Google's business, there are two very important points that we would like to cover:

- The Bavarian State Library has and retains exclusive control over the application of conservation and maintenance criteria in every process step - even at an individual case level.
- The Bavarian State Library works in close collaboration with Google to ensure that the digital scans are of the highest possible quality.

Basically, this means that the scans match the quality achieved in publicly-funded projects. Google is committed to the continuous improvement of digital scans quality and the application of its quality assurance measures. The Bavarian State Library can count itself lucky not to have been among the first group of „Google Libraries". Obviously, however, the quality of industrially mass-produced digital scans is not comparable to the quality of digital scans taken from a fifty-page incunabulum in a scanning process that takes several days to complete.
The consequence of mass digitisation is that the mass digitisation processes make it impossible or very personnel-intensive to select individual books, text sections or system groups. It is logistics which determines the digitisation processes used and their speed.
Even before the Google project commenced, around 6,000 books were removed every day from the Bavarian State Library's repository for local lending, reading room lending, inter-library lending, document delivery and official lending (mainly for digitisation projects and other maintenance measures). Naturally, this means that 6,000 books also have to be returned to the correct place in the repository every day. Although a library's service orientation isn't expressed by usage figures alone, they do to some extent reflect it. There are also around 5,000 volumes which are moved around within the repository every day for repository management optimisation purposes, and over 500 volumes arrive for storage at the repository every day. Additional daily book movements (retrievals, returns, transportation to and from the project and meta information optimisation teams) in the mass digitisation project increase the number of books movements every day to well above 20,000.

It is also worth remembering that numerous decisions have to be taken at several stages of the logistics process every day. Suitability for conservation, the completeness of meta information and its clear assignability to the book in question, as well as suitability in terms of size and the copyright situation are examples of issues which necessitate ad hoc decisions at individual book level. The project team members are prepared for this situation in training courses and intensive familiarisation phases so that they are able to make fast and accurate decisions.

The challenge facing the librarians

The challenge facing the library is far greater than the logistical challenge. The library staff have to ensure that the books are handed over to Google with clearly assignable meta information. When each production batch is handed over, the meta information pertaining to the batch in question has to be transferred with it and clearly assignable to it. Most of the meta information pertaining to the out-of-copyright books in the digitisation project is available in electronic form, though the majority was obtained through the retro-conversion of hand-written catalogues. Quite a few of the old catalogues did not contain enough information for the book's inclusion in the electronic catalogue and problematic cases could not be catalogued by autopsy, so there are data errors, some of which require urgent correction before the book in question is handed over for mass digitisation. The three most frequent errors are incomplete and incorrect conversion images, incorrect book classifications and incorrect book numbers.
Many formal errors (e.g. an incorrect book classification) can be identified by automatic catalogue database queries and printed out as error lists. Trained assistants are able to process some of these error lists after thorough training. This enables the elimination of the errors in meta information pertaining to the entire collection so that they have no influence on the provisioning of the books for the digitisation process. However, the majority of errors have to be remedied by catalogue specialists who are very familiar with the Bavarian State Library's old collection. Although the experts only making corrections where absolutely necessary, it will take quite some time to process the meta information for every book before it can be digitised. We will be discussing the consequences of this later on.

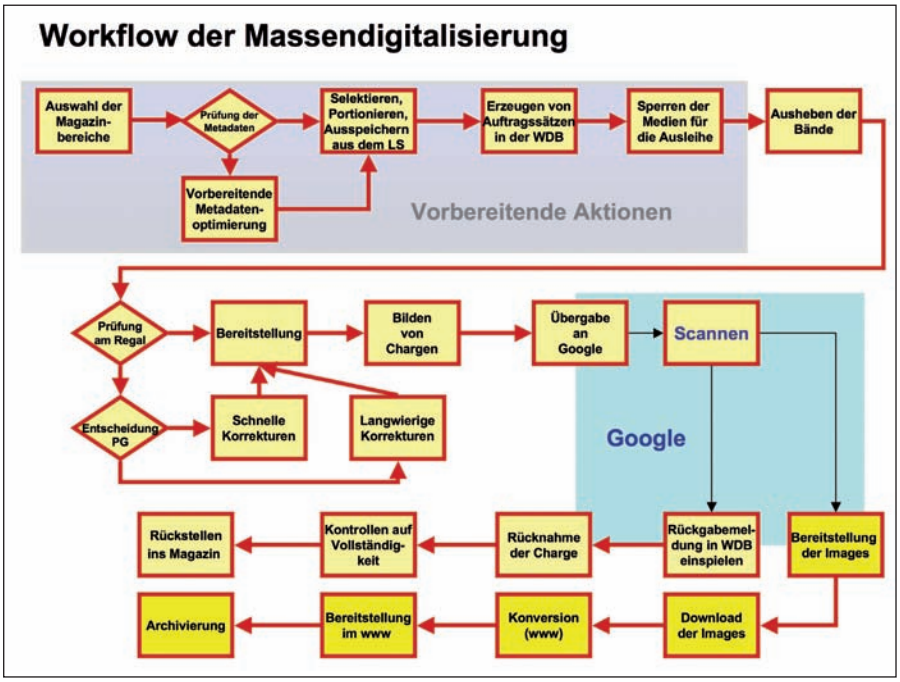

Abb. 2: Der Workflow im Überblick (LS = Lokalsystem; PG = Projektarbeitsgruppe; WDB = Workflowdatenbank)

The workflow database

We have already made good progress in the project process of "eliminating errors from the meta information". To explain the procedure, we have to provide a little more background information.
At the outset of the project, we soon realised that we needed a tool that would give us with an overview of the entire workflow

and provide us with answers to many foreseeable questions:
Is book xy with the service provider?
Which books are currently with the service provider?
Is book xy back and has it been digitised by the service provider?
Have we also received the digital library copy from book xy?
Why wasn't book xy digitised?
How long will user X have to wait for book xy?
How long will user Y from Würzburg have to wait for the digital scan to be available on the internet?
Precisely how many books have been processed to date?
Does book xy have a fold-out map or graphics that couldn't be scanned straight away?
Which books will be handed over to the service provider tomorrow morning?

It was soon clear that this instrument did not exist in the required form or only met some of our requirements. That's why we chose two software tools, ImageWare's MyBib software for library document delivery order processing and the Bavarian State Library's workflow tool and repository architecture, ZEND (Zentralen Erfassungs- und Nachweisdatenbank). We can build on these tools and both of them are effectively in use at the Bavarian State Library. ZEND and MyBib were further developed by the Bavarian State Library's Digitisation Centre in Munich and ImageWare into a tool for the management of mass digitisation projects. Both softwares were harmoniously integrated and connected via open interfaces. Together, they are called the „Bavarian State Library Workflow Database (WDB) in this publication, whereby the MyBib-based part has the provisional name MyBib-WDB and it is predominantly used for the management and documentation of logistical processes. The upgraded ZEND that is used for industrial mass digitisation is called Google-ZEND. This part controls the collection of digitised scans, the creation of WWW-compatible image formats, provisioning and long-term archiving. One key process effectively demonstrates how closely interlinked MyBib-WDB, Google-ZEND and the catalogue system are. It is the process for the catalogue verification of the digital copy. When the digital copy has been provided to the repository by Google-ZEND, automatic feedback is sent by Google-ZEND to the MyBib-WDB which initiates verification via the Bavarian Library Association's (BVB) inter-library system's catalogue and local catalogue by the entry of a persistent link at the relevant cataloguing entry. In this process, over 30 workflow steps are performed and documented by the two software tools. The most important functions are listed below:

- Guarantee of the correct assignment of the scanned object, meta information and digital copy
- Creation of digitisation order sets and print out of the appurtenant order slip
- Documentation of the reasons why a book cannot be processed
- Management of subsequent processing tasks e.g. any necessary meta information corrections
- Compilation of a production batch and documentation of all works in a batch
- Controlled hand-over of books in batches, including the appurtenant meta information, to the scanning service provider
- Acceptance of the books back from the scanning service provider and completeness checks
- Acceptance of the library digital copy back from the scanning service provider
- Publication of the digital copies on the World Wide Web
- Verification in local and national catalogues through the automatic creation of URNs and further information (resolving URL; reference to freedom from costs etc.) and their entry in the catalogues
- Transmission of image files, structural data and text data on processed works to the LRZ Leibnitz data centre's long-term archiving system

Only a very high degree of automation enables the daily processing of many hundreds of digital copies. The LRZ hosts the scalable hardware with adequate memory and the server cluster for the image conversion of the digital copies at the Bavarian State Library, and it provides the high speed data lines that are necessary to process the high daily production loads.

An overview of the workflow

The Bavarian State Library's collection of old works is categorised in just over 200 different kinds of subject classifications. The subject classification numbers were retained until 1936 and they are still used to classify all works published before this date. It therefore makes sense to use these classification numbers to select all the works which are suitable for digitisation. The sequence in which the classification numbers are processed is established many months before the digitisation process begins. The sequence is selected on the basis of projected meta information error frequency, maintenance measures to be performed or already performed, format2, location of the repository3 and the repository's mid-term relocation plan. Sections containing several 10,000s of books are selected. The meta information, particularly the book numbers, are automatically subjected to formal checks. Error lists are printed out and processed by trained assistants as explained above. The book number sections are then selected again in the local system, divided up and exported. When they have been stored in MyBib-WDB, the order sets are created. A specific number is allocated to each order set, the "digID", which is then supplemented by the book titles contained in the Bavarian State Library's local catalogue in a check back process. The digID is then the most important connecting element between the local catalogue database and the MyBib-WDB data. It is later a decisive element in the persistent link and therefore the main link between the archiving system and catalogues. It is printed as a bar code on the order slips which accompany the book through the digitisation process until it is returned to its storage place in the Bavarian State Library's repository. The digID also enables the library digital copy to find the "right place" in the Bavarian State Library's archiving system.
Around six weeks before we print the order slips in MyBib-WDB, the volumes to be digitised are taken out of circulation and loaned out books are officially reserved. By the way, if they are essential for research purposes, these books can be loaned for use in the Bavarian State Library's reading rooms until they are handed over to the service provider. The process of digitisation, our completeness checks and the return of the books takes a few weeks' time. The entire digitisation workflow makes the books unavailable for loan for around three months in total. However, they are only absolutely inaccessible for around half of this time. This is not much longer than the standard four week loan period. We would like to specifically emphasise that this is an impressive time compared with other digitisation and maintenance projects and it is an indication of quality based on thorough process planning.
The order slips are printed out and the books are retrieved from the repository on the basis of the order slips around two weeks before they are handed over to the service provider. This is a very complex process which is inadequately described by

2 The Bavarian State Library's pre-1935 works are mainly divided into the three formats of folio (≥35 cm), quarto (25–35 cm) and octavo (≤25 cm).
3 The Bavarian State Library has three repositories: the main repository in Ludwigstrasse, the repository in Garching and an alternative repository in Munich's Euroindustriepark.

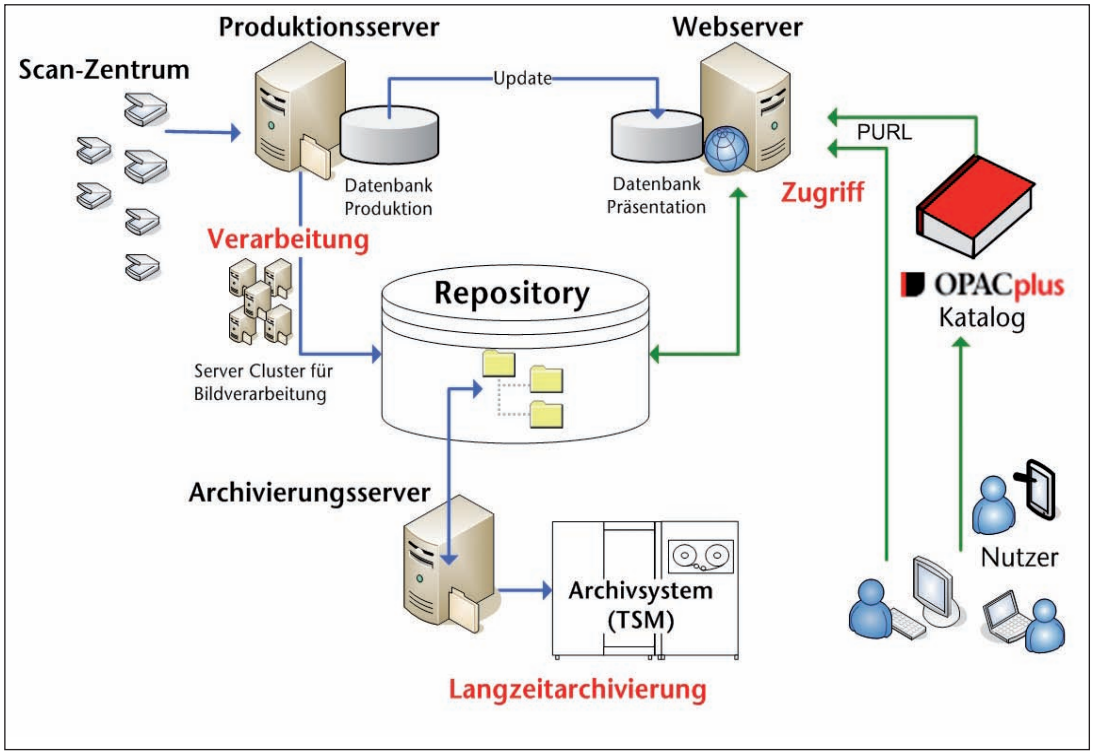

Abb. 3: Die Systemarchitektur des Bereitstellungs- und Archivierungssystems (PURL = Persistent URL; TSM = Tivoli Storage Manager)

the word "retrieve". First of all, the books have to be checked to ensure that they are suitable for scanning (condition, size, year of publication etc.). Then, one of three different things can happen:

1. The order slip is placed inside the book, which is then sent off for digitisation. The order is entered in MyBib-WDB by scanning the digID bar code with handheld scanners and integrated mobile data storage units. The digIDs are centrally input from the storage unit into the MyBib-WDB database at a later time. The individual orders are then assigned a status which permits them to be included in a batch for handing over to the scanning service provider.
2. If books are obviously unsuitable or missing – sometimes the books are put on the wrong shelf – the reason is specified on the order slip. The order slip is then scanned into MyBib-WDB to store information about the reasons why the book has not been scanned. If the situation changes or follow-up projects are implemented, it is then easier to select these works again.
3. When the person retrieving the book isn't sure whether it is suitable for scanning or discovers meta information ambiguities, the book either remains on the shelf (in which case a reprocessing note is made on the order slip) or it is shown to the local project group manager for his or her opinion.

The project group has to decide whether the books that it has received can be processed quickly within the group or whether they have to be passed on to the quality assurance and media processing teams for complex meta information corrections. Some of the books are sent to the maintenance department or the Bavarian State Library's Institute for Book and Document Restoration (IBR). All the decisions about the cultural artefacts which have been entrusted to us have to be efficient, objective and responsible. The Bavarian State Library's workflow database provides the necessary support and documentation in this process so that each book's progress can be tracked at all times.

This form of working through the repository systematically, shelf-by-shelf, is associated with the major benefit that all books for which no order slip was printed out as a result of the meta information being incorrect or incomplete can be found.

This process will provide us with two further, long-term desiderates, although this was not the original intention. All of the library's out-of-copyright and valuable old works will be subjected to a thorough audit and books which cannot be digitised as a result of serious damage will be documented and then repaired. The difference between this and other damage audits is that instead of only a small portion of the collection being audited and the result extrapolated, all books will be audited one-by-one and the findings in each case will be documented in MyBib-WDB.

The books provided to Google therefore come from three sources - books which do not require any corrections and are sent for digitisation straight away, books prepared by the local project group and a small volume of works which will come from the media processing team, a book binder or the Institute for Book and Document Restoration. MyBib-WDB then creates a "batch" consisting of a precisely defined number of books and delivers them for scanning.

So what happens to all the books with meta information that can't be corrected in time to send them for digitisation? In a few year's time, before the defined project term ends, we will be implementing the above-described process of "going through" the Bavarian State Library's entire old book repository for a second time. Obviously, all the books which have already been scanned (the majority) will not be included in the process, which will considerably reduce the time required. Now you will see why we were able to be relatively generous to users who requested book loans during the "first round". Any books that were loaned out instead of being send for digitisation will be included in the second or possibly even a third round of digitisation. This process also ensures that books which had lengthy maintenance performed on them can also be digitised.

As previously mentioned, the Google processes are not described here. The books are definitely scanned, scan quality and completeness are checked and - if possible - classification data and indexing data are generated for full text searches. Then the books are returned in complete batches to the Bavarian State Library. Batch-by-batch returns and MyBib-WDB simplify book counts to ensure that all books have actually been returned.

After their return, the books are put back in the repository by experienced staff who

can identify any missing books by gaps on the shelves. One final completeness check is made by scanning all the order slips removed from the books that were replaced on the shelves into MyBib-WDB.

Provisioning of the digital copies

The provisioning and long-term archiving of the library digital copies is organised by the Munich Digitisation Centre (MDZ) in collaboration with the Leibniz Data Centre (LRZ). An efficient technical infrastructure was set up and tested for this purpose in the months prior to project start-up. It was based on the Bavarian State Library's existing, tried-and-tested ZEND system which has been in operation at the MDZ since 2004. It was adapted for mass digitisation purposes and upgraded to include several additional functions. The central feature is a scalable repository for the storage of all project data. A central production server retrieves the data packages from the scanning service provider and transfers them to a server cluster for conversion. The data is then processed in the Munich Digitisation Centre's standard workflow. Presentation derivatives and structural data are created and a leaf-through version of each book is mapped for use. A second production server then archives the files in the TSM (Tivoli Storage Manager) system at the LRZ. The archiving system is directly connected to the production system. At the end of the entire process, the digital scan is made available on the Bavarian State Library's web server. Then, automatic feedback is provided to the WDB and, from there, to the catalogue system. Only when the catalogue link is entered is the digital scan visible and accessible to users. Finally, full text indexing makes whole book searches possible.

The system is designed so that all digitised books can be processed and provisioned soon after scanning and so that no „backlog" builds up. The entire processing chain, from data collection through image conversion to provisioning and archiving would not be possible without a high degree of automation in all production stages. Due to the vast quantity of data to be processed - several hundred books have to be processed and provisioned on every normal production day - manual processing wouldn't just slow the process down, it would also introduce errors. Fast availability for our users and the reliable archiving of all data take priority.
After comprehensive preparatory work, the BSB is well equipped to cope with the challenges of this large-scale project. In a few years' time, the majority of the Bavarian State Library's out-of-copyright historical collection will be available online. Then, it will only take users a few mouse clicks to display around one in ten of the Bavarian State Library's books on their screens at any time of day or night, from any location and without having to wait. Although these will „only" be virtual copies, all the books will be returned safely to their shelves in the BSB's repositories where they'll be available to book lovers who like the feel of the genuine article.

AUTHOR'S

Martin Baumgartner
Kooperatives Datenmanagement
Martin.Baumgartner@bsb-muenchen.de

Michael Beer
Ltg. d. Qualitätssicherung Erschließung
Michael.Beer@bsb-muenchen.de

Dr. Berthold Gillitzer
Stellv. Leitung Benutzungsdienste
Berthold.Gillitzer@bsb-muenchen.de

Dr. Wilhelm Hilpert
Leitung Benutzungsdienste
Wilhelm.Hilpert@bsb-muenchen.de

Rosmarie Leichtl
Leitung Workflowteam Google
Rosmarie.Leichtl@bsb-muenchen.de

Gabriele Messmer
Koordination Digitale Bibliothek
Gabriele.Messmer@bsb-muenchen.de

Karsten Trzcionka
Leitung Dokumentverwaltung
Karsten.Trzcionka@bsb-muenchen.de

Dr. Thomas Wolf-Klostermann
Digitale Bibliothek
Thomas.Wolf-Klostermann@
bsb-muenchen.de

Alle Autoren:
Bayerische Staatsbibliothek
80328 München