

*Digitization of books in the
National Library*
– methodology and lessons learned

National Library of Norway

September 2007

1. The digital national library

The vision of the National Library of Norway is to be a living memory bank, by being a "Multimedia Centre of Knowledge" with focus not only on preservation but also on mediation.

To succeed with this ambition, one of our main goals is to be a digital national library, as the core of a Norwegian digital library. A digital National Library is just another way of being a National Library. It is then important to have as much digital material as possible, not only historical material but also the modern part of the cultural heritage, to give access to as much material as possible to as many as possible whenever required.

The Norwegian National Library has therefore started a systematic digitisation of the entire collection. Based on a modern Legal Deposit Act we receive everything that is produced and of interest to the public, be it books, newspapers, periodicals, photos, films, music or broadcasting. All broadcasters with a license in Norway may be asked to deliver copies of their programmes for preservation, and we have an extensive collaboration with the national broadcaster NRK about preservation and dissemination. We are also to preserve those digital signals that are never converted into anything else before they reach the user. Starting in 2005 we have harvested large parts of the Norwegian web domain .no.

Some of this material is delivered to us in digital formats, and we are about to have a relatively large digital collection when it comes to the audio-visual, but little when it comes to printed material.

The issue now is how we shall take advantage of having established a digital repository for preservation in order also to give access to this material both to scholars, students and the public. This is a challenge in many aspects, but mainly with regard to copyright, technical quality and dissemination pedagogic.

Our aim to establish a unified Digital Library for all our different media, with our Digital Long Term Repository as a basis, will require the following strategies:

- As soon as possible (depending on funding, of course) to digitize all our collections. We do this systematically
- To digitize on demand
- To negotiate with the publishers to get as much material as possible deposited in digital formats
- To be a digital archive for other institutions, e.g. publishers and newspapers
- To find strategic partners to cooperate with, financially and on know-how
- To be a trusted repository for digital material in the Norwegian society
- To give access to as much as possible of our cultural heritage, e.g. through search engines
- To negotiate with the right-holders to give access to material that is not yet in the free, so that also modern books, films, and music can be searched

To be able to do this, it is also important to take part in discussions and developing the pedagogic of the net. Libraries' knowledge about the users' needs is a good experience basis on which to develop search methodology in the crossing between metadata and the methodology of search engines. Our aim is to give access to information, knowledge and experience on a given topic across media types.

2. Digitizing the Norwegian cultural heritage

Scope

For more than 10 years the National Library has been digitizing a wide range of media. The main focus for this digitization has been photographs, sound recordings and microfilm (newspapers). As a result, the digital collection today contains more than 150,000 hours of radio, more than 300,000 photographs and more than 1,000,000 newspaper pages. In addition, we have digitized more than 25,000 books over the last year.

Still, these are modest numbers compared to what we have now set out to accomplish. If we include the estimated growth of the collections through analogue legal deposit until digital legal deposit is up and running, we expect the amount of material to be digitized in the National Library digitization programme to be:

- 450,000 books
- 2,000,000 periodicals
- 4,700,000 newspapers (more than 60,000,000 pages)
- 1,300,000 pictures (photographs and postcards)
 - 60,000 posters
 - 200,000 maps
- 4,000,000 manuscripts
 - 200,000 units of sheet music
- 1,900,000 leaflets
- 1,000,000 hours of radio
 - 80,000 hours of music
- 250,000 hours of film and television

Gifts, purchases or deposited material during the program period will add to these numbers.

In order to accomplish such an ambitious program of digitization during a foreseeable period of time, the digitization activity has been greatly increased. As much as possible, we are now streamlining the process from when an object is selected for digitization until it is placed digitally in our digital long term repository, simultaneously offering web access for authorized users. With the multimedia collection of the National Library this poses special challenges, as we need to establish separate production lines for the digitization of different types of material. In addition, there are usually several variants within each type of material, which again demand different adjustments.

Putting these things in place is a challenge, both in terms of technology, logistics, organization, manpower and financing.

Legal Rights

The National Library has a legal right to digitize the collection for preservation. However, to make the collection available in the digital library, it is necessary to make agreements with the copyright holders whenever the material is still under copyright protection.

The old material with no copyright restrictions will be made available for everyone in the digital library.

The Legal Deposit Act states that every object subject to legal deposit may be made available for research and documentation. This is also valid for the digital material subject to legal deposit. However, in the digital domain there are still several unresolved challenges, e.g. privacy questions related to the interconnection of several extensive data sources (e.g. Norwegian internet pages), and the risk and consequences related to misuse through illegal copying of digital content, which are much more extensive than for paper based documents. These are the reasons why we still do not have good solutions for access to digital legal deposited material via the Internet.

When the documents are published in a traditional way but deposited in digital formats, we have to make agreements with the copyright holders to be able to give access to the digital documents. In this case we try to get agreements giving us at least the same rights that we have for material subject to legal deposit.

The National Library also works to get agreements with the copyright holders making it possible to give a broader access to the modern part of the cultural heritage. An example is the [agreement](#) between the National Library and several copyright holders organisations on giving access to books and journal articles related to "The High North". This agreement gives the National Library rights to make available approximately 1 400 complete works in the digital library. These works will be used to evaluate user behaviour and the frequency of use of the digital material. The result of the evaluation will form basis to negotiate on more permanent agreements with the copyright holders.

Financing

Preliminary estimates suggest a total cost of around 1 billion NOK for the National Library's digitization programme. Around 60% of the cost is for the digital storage, the purchase of digitization equipment and software, the development and integration of systems that will be part of the process of digitization and post-processing, in addition to wages for carrying out the digitization and money for some external commissions. The remaining 40% of the cost will be needed for the indexing of the material in order to establish the metadata required for retrieval, and for fetching the material from the collections, for the necessary conservation, and for the return of the material to the collections.

Through a re-channelling of activities towards the strategic initiatives, the National Library has prioritized 12 M NOK a year for the digitization programme. In addition, we have received a grant from the Ministry of 3 M NOK for digitization in 2007. Finally, on top of this, we have a budget of 13 M NOK for the purchase of digital storage in 2007, so the total budget for the initiative is 28 M NOK this year.

Preliminary estimates show that the whole digitization programme can be carried out in 15 - 20 years. Naturally, there is considerable uncertainty associated with an extensive and long-term effort such as this. We expect developments in technology both in terms of digital storage, digitization equipment and of tools that will enable retrieval in a way that reduces the need for manual indexing. However, it is difficult to accurately predict the consequences of such a development, in terms of both cost and increased efficiency.

Status

Large parts of 2006 were spent on carrying out the process of inviting tenders for the purchase of digitization equipment, software, digital storage and other ICT solutions required. In addition, a project manager and some system developers were hired late 2006.

In the new digitization programme, books were given priority. A closer scrutiny of this work can be found in a separate chapter.

Following an invitation for tenders we have outsourced the digitization of microfilm to a German company (CCS). So far 400,000 pages from the newspapers *Aftenposten* and *Adresseavisen* have been digitized. However, this number will be increased in the fall of 2007.

In addition, the digitization of photos has been made more efficient through the purchase of software for automatic post processing of digitized material, and through an upgrade of the digital cameras used for this work. We have also bought efficient equipment for the digitization of 35mm roll film negatives.

We are also working on an upgrade of equipment for the digitization of music, radio and moving images. An important part of this is a more automated connection between the digitized material and metadata that describes the material.

Other types of material will be digitized on demand, but so far we have not established adequate production lines that will enable speedy transfer of digitized materials into our long term digital repository, and immediate availability in our digital library.

The systematic digitization is the foundation of the digitizing process. In addition, we will carry out on-demand digitization which can be both user initiated and based on the strategic priorities of the National Library. On-demand digitization will be given priority over the systematic work. This applies to all parts of the digitization process. However, at this time it is only true for those types of material for which production lines have already been established (books, photography and sound).

Future plans

Of most importance in the short term is to make operative all the production lines for the digitization of books. What remains to be done is mainly to establish a regime for quality assurance of the digitized books, and an adequate handling of exceptions through all of the production line. This includes all stages of the process starting with ordering material out from the stacks and ending with the digital versions of the material residing in the digital long term repository and becoming accessible via the NB digital library.

We have also started work on establishing production lines to preserve both digitized newspaper pages and digitally deposited newspapers in the digital long term repository, and to make them available in our digital library. In addition to the newspapers that have been digitized from microfilm, we have established a pilot of daily legal deposit from two leading national newspapers in preservation quality PDF format. When this pilot gets into regular operations, we will work towards extending it to more newspaper titles.

We also plan to start the work of establishing a production line for the digitization of manuscripts. This is a type of material in heavy demand from our users and which will be of relevance in connection with upcoming authors' anniversaries in the years ahead. For this material, in addition to digitization, we must establish the necessary metadata in order to achieve satisfactory search and retrieval in the digital library.

Later again, we will establish production lines for the digitization of periodicals and posters. In addition, we plan to automate parts of the existing production line for the digitization of photography, in order to increase efficiency.

Another very important activity will be to initiate legal deposit of new publications in digital formats of preservation quality. Besides the pilot setup with daily legal deposit from two newspapers, today we have an operative solution for the digital deposit of radio in preservation quality. Also, audio books produced in digital format are delivered from the Norwegian Library of Talking Books and Braille.

We also have an ongoing dialogue with publishers and TV broadcasters, planning to start digital deposit of books, periodicals and television during the coming year.

3. Establishing a production line for the digitization of books

3.1 Starting out

In establishing a brand new process for the digitization of books, it was important to achieve a well integrated production line capable of covering all the steps through which the book would pass before it was preserved in the National Library's digital long term repository, and also becoming available to authorized users in the digital library. We wanted to automate as much as possible of the data flow and processing of the digital book, at the same time leaving enough flexibility in order to accommodate new process stages in the production line if the need should arise.

The processes we wanted to include were: selection for digitization and ordering of the books in question, fetching of material from the stacks, transport, extraction of metadata from the catalogue, digitization, OCR treatment and structural analysis, format conversion, the generation of preservation objects, ingest to the digital long term repository, notifying the catalogue of the digital object, and indexing OCR text and metadata in our search engine.

We also saw that different types of scanner technology gave very different efficiency in the digitization. If books could be dismounted, the scanning would become at least 10 times faster.

3.2 Fundamental decisions

Preservation – quality and format decisions

The digitization programme is a part of the National Library's strategy for preservation and dissemination. Digitization will make preservation of the collection more efficient and less vulnerable to physical deterioration. This means that the digitization must be performed in a quality so high that, after digital preservation, it must be possible to satisfactorily recreate the properties of the original at a later time. At the same time, the digitization and the chosen formats must fulfil the requirements for dissemination of the material.

We have chosen to digitize books at a resolution of 400 dpi and a colour depth of 24 bits. Our preservation format is JPEG2000 with lossless compression. The preserved image is not processed or reduced in any way after the scanning process.

By choosing losslessly compressed JPEG2000 instead of uncompressed TIFF as a preservation format, we will reduce the need for digital storage by about 50%. For the whole digitization programme this means savings on the order of NOK 70 M. Through practical tests we have demonstrated that we are able to convert from the JPEG2000 format back to uncompressed TIFF with absolutely no loss of information.

An argument against using JPEG2000 is that one bit error in a JPEG2000 image file will be able to destroy the whole image, whereas a bit error in an uncompressed TIFF image file will not affect more than one pixel. With the storage policy in our digital long term repository, we find the risk of bit errors to be negligible.

The quality requirements for preservation are far stricter than those normally used for dissemination. Also, the formats used for dissemination have a shorter lifespan, partially because new formats are developed with more advanced compression algorithms that offer better quality with less data, and partially because there are new versions developed of the existing formats, with better algorithms and better quality. Therefore, we have chosen to generate the dissemination format from the preservation file at the moment a user asks for the image. Using this strategy we will easily be able to switch dissemination formats by replacing the algorithm that generates the dissemination format.

In today's solution we generate a JPEG file of the desired quality for viewing (typically around 200 Kbytes) from the JPEG2000 file located in the digital long term repository (typically around 20 Mbytes).

Perform our own digitization or hire others?

Today, there are several organizations offering cultural institutions to digitize their book collections for free or at a very low price, in return asking for the right to store the digital copies and to offer search and display of the books. Examples of this are Google and the Internet Archive. In the National Library's book collection there are relatively few books which are no longer protected by copyright, and to which free access accordingly can be offered. Out of the 450,000 titles that will be digitized under the digitization programme, at present only around 5,000 titles are no longer copyrighted. It is important to the National Library that books still under copyright will only be stored in the National Library's digital long term repository. Access to such books will only be given in accordance with agreements made with the right holders. Also, it has been a fundamental principle for the National Library that other service providers must be given equal opportunity to offer services based on our collections. The sum of these considerations has meant that we have not deemed it natural to cooperate with this type of organizations for digitization. Still, we will invite them to disseminate what we have digitized, on par with other actors.

It has also been part of the picture that the National Library has been able to reassign existing staff to tasks associated with the digitization production line. This has meant that the total cost of in-house digitization of books is lower than it would have been for outsourcing the digitization.

For other types of material we have chosen to outsource digitization. For instance, this applies to digitization, OCR and structural analysis of newspapers on microfilm.

Dismounting books

In order to achieve efficient digitization, we have chosen to dismount books for digitization if we have at least three copies in our repository library. The dismounted copy is then thrown out after digitization.

When we have fewer copies, the books are scanned manually, with operators opening the books and scanning two pages at a time. The most vulnerable books are scanned under the supervision of a conservator, and any necessary conservation measures are performed before or in connection with digitization.

The process of preparing books for dismounted scanning, is more labour intensive than preparing books for manual scanning. Special operators are required for deconstructing the books (separating the binding from the rest of the book, removing glue with a hydraulic cutter), and the scanning of the binding is a separate process. Because of this, 4 operators are needed in order to feed one scanner of dismounted books. Still, the total picture means lower cost and higher production than if the same resources had been used for manual scanning.

As of today, around a quarter of the National Library's book collection can be dismounted for digitization. In order to improve this ratio, the National Library plans to invite Norwegian libraries to contribute copies of books we have too few of. In this way, book digitization can be carried out faster and more efficiently in the National Library, and the libraries will be able to free up space in their stacks.

We have not yet tried out scanners that automatically turn the pages of whole books that have not been dismounted. Such scanners are developing rapidly, and this technology is becoming very interesting. For books that can not be dismounted this technology may offer considerably higher production per operator, both because the scanners are fast and treat the material gently, and because one operator can serve several such scanners at the same time. However, investment costs for this type of scanner are still high.

OCR and structural analysis

In order to allow for full text search, all digitized books undergo a process of optical character recognition (OCR). In the regular production this process is fully automated, and there is no manual quality control or correction phase. The text that results from the OCR treatment is indexed in our search engine together with the metadata. If a search gives results from the text, the page of the book where the text was found will be displayed, and the user can browse the book from that page.

Also, an automated structural analysis is performed, during which any table of contents is annotated, and page numbers in the book are verified so that the user interface relates to the actual pagination of the book. This is also an automatic process. The software allows for very advanced structural analysis, but at this stage it is not feasible to increase complexity without also applying extensive manual quality control and post control. For selected parts of the collection we will perform more advanced structural analysis including annotation of several parts of the documents, again allowing for more advanced navigation of the books in the user interface.

At present, about 2,000 – 3,000 books are digitized every month in the National Library. With this volume it is not possible in practice to perform manual post control of the OCR and structural treatment.

Both OCR and structural analysis are performed using software called docWorks.

The National Library's digital long term repository

The digital long term repository is an infrastructure for the long term preservation of digital objects. Everything that is digitized as part of the National Library's digitization programme is to be preserved as digital objects in the National Library's digital long term repository.

The digital long term repository separates the use of digital content from the technology which is employed for the storage. This allows for easy migration to new generations of storage technology without affecting the systems for retrieval of the digital content. This is very important in a 1000 year perspective.

All digital content is stored in three copies on two separate storage media in the digital long term repository. At present one copy is stored on disk while two are on tape.

Search engine

In order to realize search across large aggregations of data, the National Library has chosen to employ search engine technology rather than traditional data base solutions. Both metadata and full text are indexed by the search engine, and searches are performed without regard to types of material. We have also implemented a so-called drill-down search in the metadata. Metadata for objects satisfying the search criteria is analyzed in real time during searches, and alternate paths of navigation and different ways of narrowing the search results are built and displayed to the user.

The search engine used is delivered by FAST.

Authentication and authorization – access control

The National Library has chosen to employ role based access control. We have chosen to cooperate with FEIDE, which is the national infrastructure for authentication and authorization of users at Norwegian universities and other institutions of higher education. This means that we can open for access to defined groups of scientists at the universities and institutions, without knowing every individual. The universities are responsible for the authentication of those who satisfy the requirements for the different roles which have been defined within the system.

If we had picked an access control based on user names and passwords with associated rights for individuals, the National Library would have needed to spend considerable resources on the administration of users and the maintenance of access rights.

At present we have not implemented this access control solution for digital books. Therefore, we have only indexed and made available books that are either out of copyright, or covered by agreements with the right-holders on giving open access.

The software used for user authorization and authentication was supplied by SUN Microsystems.

3.3 Production lines

Prioritization

The basis for the digitization is the systematic selection. We have chosen to start with the oldest material in order to quickly get the material which is out of copyright into our digital library. In addition to the systematic selection we give priority to material on the basis of internal needs and external requests. Especially prioritized material is placed at the head of the queue, in front of the systematic selection.

A special case is the agreement between the National Library and several rights organizations regarding books and articles related to the "High North". These works have been given special priority in the digitization process.

Ordering and extraction from the stacks

In order to achieve an efficient extraction of material for digitization, a special function has been implemented in BIBSYS, our book cataloguing system. Here we can order a given number of titles for dismounting, chosen automatically by the system among titles that there are a sufficient number

of copies of, starting with the oldest. In addition, we can order single titles to be given special priority (both on extraction from the stacks and through the whole production line). Adaptations have also been made to the software that runs our automatic storage system for books, so that the operators can give top priority to remote loans, and then extract books for digitization. This system is integrated with the catalogue, so that books ordered for digitization automatically appear in the interface for the operators of the automatic storage system.

Already we have spent more than one man-year on system adaptations of the catalogue system and the software for the automatic storage system.

Digitization

For the books to be dismantled we have two hydraulic cutters, three binding scanners (i2s Copibook) and two auto-feed scanners (Agfa S655). For the page-turner scanning we use i2s Digibook Suprascan. Five of these are A2 scanners for normal page-turner scanning and one is an A0 scanner for special material. The A0 scanner is operated by conservators.

Before the bindings are scanned, all metadata for the book is retrieved from the catalogue (BIBSYS) by way of a bar code assigned to every book in BIBSYS. A digital ID for the book is generated and inserted into an XML file together with the metadata obtained from the catalogue.

In the case of autoscan, a sheet of paper with the bar code of the digital ID is printed out after the scanning of the binding. This sheet is put on top of the stack of the dismantled book. When the bar code later is sent through the autoscanner, it is identified, thus making an automatic connection between the metadata file and the scanned binding.

In the case of page-turner scanning, the binding and the content of the book are scanned on the same equipment. This process too fetches metadata from the catalogue and generates an XML file with metadata that accompanies the book through the rest of the process.

OCR/DSA

After digitization the digital book and its accompanying metadata will be placed on temporary storage ready for further processing. The books must be manually imported into the docWorks software, but from there on the processing of most books is fully automated. Manual operators are only employed for handling of exceptions when the software calls attention to errors in processing (i.e. when processing failed to stay within the defined tolerance limits).

In addition, operators are used for quality control of special parts of the collection that we want to process further.

Books of high priority are placed in special folders that are imported before the standard systematic digitization.

After OCR treatment and document structural analysis, losslessly compressed JPEG2000 files are generated from all the image files of the book. This is the format used for preservation.

Digital preservation

After processing in docWorks, a METS object containing metadata, the digital book, the OCR processed text and structural information is generated. This object is placed in the National Library's digital long term repository for preservation.

Simultaneously the catalogue is updated with the digital ID of the book.

Indexing

At regular intervals, an OAI import of catalogue data is performed. If this import finds that a book has been updated with a digital ID, a process is started in order to fetch the metadata and text of the book from the digital long term repository and index both, making the book available for search in the digital national library.

4. Important lessons learned this far

Scope, complexity and implementation

Implementing an integrated production line for the digitization of books with a high degree of automation turned out to be far more extensive and complicated than anticipated.

In order not to waste time, as soon as the decision to establish a production line for the digitization of books had been made, we launched activities aimed at realizing the first part of the chain of production (ordering, extraction, transportation and the digitization itself). In order to realize the efficient extraction of material, adaptations had to be made both to BIBSYS, which is the National Library's book cataloguing system, and to the software that runs the automated storage system from which the books are retrieved. This made us dependent on two external suppliers, which had consequences for the rate of development.

Carrying out an invitation for tenders for the purchase of scanners is also a long and time consuming process. We could not develop the method of digitization until it was clear what kind of equipment we would be using, and then we had to get the first scanners installed before we were able to terminate the implementation and start testing.

After the first part of the production line was in place, we began test production. Having aimed at a high production rate, we soon found ourselves with large amounts of data in temporary storage. While waiting for the rest of the production line to be put in place, we had to establish temporary routines for the safeguarding of the digital content.

In order to set up the rest of the production line and the functionality of our digital library that would facilitate the search and display of books, a multitude of development activities had to be launched. Some examples: Installing and putting into operation the software for OCR and structural analysis of documents, and the integration of this system into the production line, the generation of preservation objects based on the METS standard, the process of ingesting the METS objects into the digital long term repository, a setup for updating the catalogue system with the digital ID, OAI harvesting of metadata from the catalogue system, the process which retrieves text and metadata from the digital long term repository for books which have a digital ID, the indexing of these, and the development of necessary functionality for the search and retrieval of books in the digital library. At the same time there arose great pressure to quickly get to see the results of the digitization already under way, which led to the pushing of deadlines and, during one period, a very high level of stress among the section responsible for development.

After the functionality for the display of books in the digital library was in place, it soon became clear that this would become a very interesting service which would provide a "lift" to our digital library. Also, we had received considerable media coverage of the digitization, and expectations of seeing results were great, both from external users and from inside the National Library. We

therefore decided to launch the service, even though the production line still was under development, meaning that many manual operations were needed to speed a book through the full production line. The service has functioned well, but the expectations of quickly reaching a large volume of digital books in the service were not fulfilled. This was mainly due to the fact that the production line was not fully implemented, and accordingly not put into regular service.

With hindsight it is easy to see that we ought to have had a greater focus from the beginning on the unity of both the whole production line and the necessary functionality in the digital library, and that from the start we should have better assigned ownership to the timeline in this development work in the whole organization.

Actual efficiency

With basis in the specifications of the digitizing equipment, we assigned production targets already from the start. These took into consideration that this was a ramp-up phase. Still, it turned out that some factors we had not been aware of reduced the overall efficiency. This became most apparent when looking at the automated scanners.

The book pages were on the whole thicker than the reference paper used to measure the scanners' specifications. This led to a decrease in paper feed speed, which made a great impact on the daily production numbers.

Since we started with the oldest books, we had an issue with them being very dusty. This meant that the scanners needed considerably more cleaning and maintenance than expected by the supplier. This in turn meant reduced production time per day, and therefore reduced daily production compared to expectations.

Our plan called for the scanners to work as continuously as possible through the whole working day. This was to be realized through operators relieving each other at the scanning stations, taking their breaks at different times. This was a new and unfamiliar way of working, creating some resistance among the operators. In practice we have not been able to realize this well enough, and also this has contributed to reduced production time per day on the scanners compared to our prognoses.

Actual production has been between 60% and 80% of our stated production goals.

Quality

During initial testing, all pages of the books digitized were subjected to quality control. Since then we have not had routines for quality assurance of the digitization work. After the functionality of the digital library became capable of displaying the books, there have been random quality checks of the quality of what is available through the service.

Books that were displayed in the digital library revealed that the compression of the digital display copies had been performed using inadequate settings. The visual quality of the images was lower than expected. After adjustments had been made, the results became much better.

Obviously, the quality of the digital books is closely related to the quality of the originals, Our algorithm for the automatic selection of books for digitization does not take this into account, and we accordingly run the risk of extracting inferior specimens for digitization even if there actually are very good copies in the collection.

The scanners that digitize the dismantled books scan both sides of every sheet in one operation. This means that two different digitizing units are processing the two sides of a sheet. It has proved quite difficult to calibrate these two units identically, resulting in colour variations between the pages. This has improved immensely since the test phase, but the problem has not yet been fully solved. Experiments have been made with scanning a reference sheet at the start of every book, to facilitate subsequent automatic colour adjustment. So far these experiments have not yielded the desired results, but we will strive towards a solution.

When the production line enters a phase of regular production, we plan to establish random quality control of digitized books.

OCR/DSA

From the outset, we had planned for fully automated use of OCR and document analysis tools. These kinds of tasks had never before been performed at such a large scale, and we had no prior experience using tools of such advanced nature.

The first challenge was to establish a large scale production setup with eight instances of the software on eight blade servers. This was necessary in order to achieve sufficient processing capacity, but it turned out to be harder than expected to make this stable and operative.

The next challenge arose when it turned out to be impossible to run the system fully automated. This created an unforeseen need for resources, and manning this task caused us some headaches. We spent some time figuring this out, and had to postpone the planned training on the system. This in turn gave us a short-term competence problem since it is a complex system with very advanced functionality. The answer was in part found through close contact with the supplier. This challenge has now been overcome, and training has taken place.

Our initial expectations of precision in the fully automated structural analysis have so far not been met. Advanced structural analysis can be done, but with such a degree of uncertainty that manual quality control is absolutely necessary. The more advanced the analysis you want to employ, the more manual quality control you will need. Accordingly, we will use this only as an exception, in special dissemination projects. A simple calculation shows that a post control that takes on average 15 seconds per page, will in total require an effort equal to 18 man days per day at the present production level. For this we don't have resources. So we have been forced to stay at an absolute minimum level by only requiring correct pagination in the digital library service and that the table of contents must be directly linkable whenever a table of content is present.

So far we have focused on OCR and DSA of publications in Latin letters. Here we have acceptable precision in the letter recognition. For Gothic letters the results are worse, but even there we have a degree of recognition that opens up interesting possibilities for free text search. We are running separate configurations of the system for Latin and Gothic letters. Books are categorized when the bindings are scanned, and then they are routed to the correct configuration. We expect there to be a potential for improvement by further training of the software and more advanced configuration of the system.

The digital long term repository – scaling and performance

So far we have operated a separate instance of the DSM (digital long term repository) for the digitization of books. The use so far does not show performance problems for the DSM, but the usage is expected to rise considerably compared to the present traffic, as the volume of digital books offered in the service increases.

Since we do not have in place an access solution (user authorization and authentication) for books, we have so far chosen to only place in the DSM books that we are allowed to give full access to in the digital library. Some of the logic today is that when a book is placed in the DSM its digital ID is entered into the catalogue. This in turn lets the book be fetched automatically from the DSM for indexing to the search index of our digital library.

This strategy means that most of the digitized books are still in temporary storage, albeit subject to the same storage policy as those in the DSM.

When a user asks for a certain book page at a certain quality (at present the interface gives a choice of three quality levels), a JPEG file of the desired quality is generated from the JPEG2000 file in the DSM. So far we have not implemented intelligent pre-caching or placing pages in a buffer outside the DSM to increase performance as perceived by the user, but this may be an interesting future development.

So far the rate of progress in technology has given us enough flexibility that we can operate a one machine solution for the digital long term repository (DSM). But this makes it vulnerable when errors arise. We are therefore continually evaluating the possibility for more robust solutions.

Statistical tools – production supervision tools

So far we have been using simple UNIX tools to generate the statistics necessary in order to supervise production. We are now considering the development of more advanced general production supervision tools that will at any time give us updated information of where a given object is in the process. We should also be able to use them to generate production statistics from all stages of production.

Exception handling

With a few exceptions everything in the production line that can be automated, has been automated. Still, some times there can arise deviations in all stages of the production line, and these deviations must be handled and followed up manually. This has been one of the greatest challenges in the production so far. We are now working on establishing routines and assigning precise responsibility for such follow-up in the line organization.

5. Summary

The implementation of a production line for books has had its fair share of problems. However, we have learned from our mistakes, and today we have in place an advanced production line for books. By the end of the year the production line will come into regular operation, and the last pieces of assigning responsibility and exception handling will fall into place.

In spite of the challenges we have met on the way, we have had a considerable production during our test year. Close to 26,000 books with an average of more than 200 pages per book have now been digitized, and most of these have also undergone OCR and structural analysis. A little more than 1,500 books are freely accessible in their entirety in our digital library, where they are also searchable in full text.

The challenge we now face is to establish production lines for all the types of material that are to be digitized, so that we truly will be able to establish the multimedia digital national library.