LIBER QUARTERLY 18 (1), April 2008 – ISSN: 1435-5205. P28-38
http://liber.library.uu.nl/
Igitur, Utrecht Publishing & Archiving Services

# Mass Digitisation by Libraries: Issues concerning Organisation, Quality and Efficiency

**Astrid Verheusen**

Research & Development Department, Koninklijke Bibliotheek,
PO Box 90407, 2509 LK The Hague, Netherlands, astrid.verheusen@kb.nl

Ever since the world-wide web made it possible to display graphics on the Internet, libraries have been scanning their older documents and pictures to provide access to them. From the middle of the 1990s thousands of libraries of all sizes began scanning parts of their collections, provided these with metadata and made them available on the web. The emphasis in these first, rather small, digitisation projects was on experimenting with different techniques for both scanning and building interfaces for the Internet. Along the way, methods for quality assurance, project management and business models became more professional. In line with the progress made in the field of digitisation, a profound knowledge of best practices has been developed. However, this knowledge is not available for all cultural heritage institutions who want to digitise their collections. Most of the smaller institutions lack experience and, moreover, the means to digitise in an efficient way. At the same time, the larger libraries are moving towards large-scale digitisation of historical texts while Google has already digitised millions of books from several libraries around the world. Although many libraries welcome the unprecedented access to all this information, Google has also been criticized for the inferior quality of their images, the emphasis on the English language, the violation of copyright laws and the lack of attention for preservation issues. The question therefore arises: can libraries do better than Google?

With the i2010 vision of a European Digital Library, the EU has launched an ambitious plan for large-scale digitisation projects aimed at transforming Europe’s printed heritage into digitally available resources. However, lack of knowledge and expertise slows down the pace with which this vision can be realised. What, then, are the main obstacles to large-scale digitisation, and how can we speed up the process and disseminate knowledge? In this paper the experiences from a number of large-scale digitisation projects at the Koninklijke Bibliotheek, the National Library of the Netherlands, will be reviewed to show how several aspects of the digitisation process can improved in order to achieve more efficiency in large-scale digitisation.

The Koninklijke Bibliotheek is in charge of several large-scale digitisation projects. In the coming years millions of pages from newspapers, magazines and books will be digitised and made available on the web. To be able to do this, the Koninklijke Bibliotheek has had to change the way in which digitisation projects are managed; at the same time it is in the process of rethinking quality standards. The entire workflow for digitisation projects is being adjusted to allow for mass digitisation. This includes the selection process, copyright issues, preparation of the materials, cooperation with the publishing sector, the specifications for the digitisation itself, research into improvement of optical character recognition (OCR), research into several file formats to reduce the cost of storage, automatic quality control mechanisms, new language-based techniques for search and retrieval, the digital preservation of the files and the technical infrastructure to support all this.

## Mass Digitisation

First of all, questions arise as to the concept of 'mass digitisation'. What is mass digitisation about? Is it about selection, or rather about no selection at all? Is it about copyright issues, finance, OCR or the quality of scanning? Who is doing it and who isn't? Is the Koninklijke Bibliotheek digitising *en masse* because of the size of its projects? Is it a question of numbers? What is the difference between mass digitisation and digitisation on a large-scale?

While there is a lot going on in mass digitisation, a clear definition of the term is still lacking. When you Google the words 'mass digitisation' you will not get an answer either. Talking of Google: there is no doubt that the Google books programme constitutes mass digitisation. Let us derive some possible characteristics of mass digitisation from Google's programme: mass digitisation seems to be about millions of books rather than millions of pages. There does not seem to be a real selection process. Until recently there was a focus on the English language, although Spanish, German and even Dutch books are now being digitised by Google as well. The books seem to be 'selected' only insofar as they should be in a format with which the Google scanning technology can cope. Mass digitisation is also about books, or rather text: it is not about pictures and also not about the special collections that many libraries hold, like maps or manuscripts.

Although most libraries seem to be in favour of mass digitisation, there is also a lot of criticism of Google's digitisation efforts. The images are of inferior quality and copyright and long-term preservation issues are being ignored.

## Digitisation at the Koninklijke Bibliotheek

Digitisation at the Koninklijke Bibliotheek seems very different from what is known about the Google books programme. To understand the way in which digitisation is taking place at the present time, one has to go back some twelve years, to when libraries all over the world started digitising their collections. In those early years only highlights of collections were digitised and these were mainly used to build attractive exhibitions on the Internet. The digitisation projects were small and their focus was on images of visually attractive source materials. There was a lot of emphasis on technical issues: how to create a good image and how to make these images available and searchable on the web. Libraries started to think about new possibilities and they also began to cooperate in digitisation projects on a small scale. In those early years the Koninklijke Bibliotheek designed several web exhibitions with highlights from the special collections, like maps, manuscripts and old sheet music.[1]

At the turn of the century a shift in emphasis occurred in digitisation activities. The Koninklijke Bibliotheek moved from digitising highlights to digitising complete collections. Digitisation projects became larger and therefore project management became a more important issue. Developments in methods and techniques were stabilizing and there was a growing awareness of the problem of long-term preservation of the digital files. Instead of visually attractive materials the Koninklijke Bibliotheek started digitising text materials and audio and video collections. New possibilities for the use of the digitised collections were discovered, such as applications for specific target groups like scientists and students.

During this period the Memory of the Netherlands was developed. The Memory of the Netherlands is the national digitisation programme of the

Koninklijke Bibliotheek that coordinates the digitisation activities of about fifty cultural heritage institutions in the Netherlands. Their collections were made available on the website, which now contains some 350,000 objects. Most of these objects are images; there are hardly any text sources available on the website.

So much for the past. The present strategic plan of the Koninklijke Bibliotheek foresees the development of a national programme for the mass digitisation of information resources in the 'humanities'. This programme should facilitate new research by scientists but also aims at the public at large. Besides digitisation, the Koninklijke Bibliotheek is working on the development of services and standards and it includes a special department involved with research and development in the area of digital preservation.

A major shift also took place in the type of material that is being digitised. The library almost completely turned to text digitisation instead of creating images of visual materials. This causes new challenges in the development of techniques for navigation, OCR, digitisation from microfilm, and search and retrieval. Microfilm as a medium for conservation purposes was abandoned. In its stead, guidelines were developed for preservation imaging within the national preservation programme Metamorfoze.

However, the main difference between the current projects and those executed ten years ago lies in the size of the projects. Table 1 contains an overview of current digitisation projects at the Koninklijke Bibliotheek. Although

*Table 1: Digitisation projects at the Koninklijke Bibliotheek, 2007–2011.*

| **Project** | **Number of pages** |
|---|---:|
| Dutch parliamentary papers 1814–1995 | 2,300,000 |
| Dutch daily newspapers 1618–1995 | 8,000,000 |
| Special collections – books before 1800 | 1,300,000 |
| Radio news bulletins | 1,500,000 |
| Metamorfoze – preservation imaging[2] | 28,000,000 |
| Books from Aceh | 200,000 |
| Memory of the Netherlands | 350,000 |
| **Total** | **41,650,000** |

the figures look very impressive, these projects do not really constitute 'mass' digitisation. Only a selection of materials is being digitised and special collections such as old books before 1800 and old newspapers are included. Copyright issues are being taken seriously, there are very high quality standards for imaging and OCR, and long-term preservation of the digital files is part of each project. But the Koninklijke Bibliotheek is certainly digitising on a very large-scale and ideas for even more projects are under way. Within the next four or five years, the Koninklijke Bibliotheek is going to digitise at least 40 million pages while planning even larger projects. However, there are some problems to be overcome before getting there.

## Obstacles to Mass Digitisation

First of all there is the matter of costs. The average cost of digitising one page amounts to € 1.30 (KB project average). This includes everything: digitisation, the addition of metadata, OCR, person months for selection, preparation of materials and quality assurance, overhead costs and hardware and software. Even though the costs for digitisation have declined during the last ten years, digitisation is still rather expensive. At the same time both the exploitation costs for maintaining websites and the costs for storage after a project has ended are increasing rapidly. Especially the costs of long-term preservation of master files are rising very rapidly as the amount of files to be preserved increases with millions each year. It is therefore necessary to reduce the costs for digitisation in order to be able to digitise more pages within the available budget.

The technical infrastructure also causes problems. 40 million digitised pages require more than a Petabyte of storage capacity, which is not easily realised. From this year onwards about 2 million files a month have to be processed. These files need to go through a quality assurance system and subsequently they need to be indexed and stored in different storage systems. In the end they should be made available on the web. At the moment the infrastructure of the Koninklijke Bibliotheek is not capable of dealing with such large amounts of files. Furthermore, the search and retrieval system is not capable of giving the end user what he needs to find. All in all, the digitisation process is very slow: in order facilitate digitisation on a large-scale, the pace should quicken considerably.

Libraries have been involved in digitisation for many years. Why is it so hard then to keep up with companies like Google who only just began digitising a few years ago? Is it only a matter of insufficient budgets? In March 2006 a symposium about mass digitisation was held at the University of Michigan. The report of the symposium included a statement about one of the obstacles for mass digitisation. It states: 'We cannot slow down to make things perfect. The rising tide will lift all boats.'[3] The idea behind this statement is that it is not only a lack of funding which makes it so difficult to digitise on a large-scale. Digitisation projects should become more efficient to speed up the process of digitisation.

At the Koninklijke Bibliotheek digitisation activities were evaluated in 2007. All steps in the digitisation workflow were reviewed. Since then things slowly began to change. This change is still in its first phase but for many aspects of the workflow it is already possible to see the effect of new approaches. A comparison between old and new approaches for some of the most important aspects of the workflow is given below.

## Improving Digitisation Workflow

### Selection and Preparation of Materials

The first step in each digitisation project usually consists of selection and subsequently the preparation of the source material for digitisation. At the Koninklijke Bibliotheek much attention is paid to both. For several digitisation projects advisory commissions have been created, consisting of subject specialists who advise on the selection of materials. In most of the digitisation projects much time is also devoted to completing the collections. For example, twenty different sets of Parliamentary Papers (about 2,3 million pages) were acquired to build one complete set for microfilming, digitisation and conservation purposes. Project staff turned every single page looking for missing pages that were completed with pages from other sets. A large team worked on this for several years and the process slowed down the project considerably.

Not only are collections completed, pages that are in bad shape will be replaced or repaired if possible. Due to the high quality standards for selection and preparation of materials, this process takes up about 10 to 15 percent of the available budget for digitisation projects.

To lower the costs and increase the pace of digitisation it is necessary to limit the time spent on selection and preparation of materials. Because of limited budgets selection is necessary but it is more a matter of priority – what to digitise first –, considering that in the end all important collections will be digitised anyway. One should at least try to digitise complete collections and not select objects from within a collection.

The time spent on the preparation of materials can also be limited. This will have consequences for the final quality of the digital files, but a comparative assessment has to be made to balance between reasonable quality and perfect solutions. In the newspaper project of the Koninklijke Bibliotheek a decision was therefore made to put limited effort into completion and restoration of the source materials.

**Digitisation**

When the Koninklijke Bibliotheek started its digitisation activities, a lot of effort was put into creating the perfect image. Also, scanning had to be carried out in such a way that the original was not damaged. This slowed down the digitisation process for the more vulnerable library materials considerably. Two or three different copies of one image were always delivered, as most libraries used to do: a master file for long-term preservation and reuse in the future, and access copies for presentation on the internet. For master files mostly the TIFF format was used.

In the early years the Koninklijke Bibliotheek bought its own scanners and hired its own staff, but it was soon discovered that scanning is not the core business of the library. From then on digitisation was outsourced. Because high standards were set, just a few companies in the Netherlands could meet the KB's requirements. The high standards led to very high quality images, but also produced very high costs for scanning and storage of the large master files.

Because of the growing scale of digitisation projects some of these old basic assumptions have to be reconsidered. There is now a preference for digitising from microfilm which mostly delivers lower quality images but is cheaper than digitising from originals. Also, research is being done into alternative file formats for TIFF like jPeg2000, which decreases the amount of necessary storage capacity. For the same reason the use of only one format for both access and preservation is being considered. For the Koninklijke Bibliotheek to be able to outsource its scanning activities, it is necessary to introduce quality standards that commercial companies can handle. While outsourcing all scanning activities, expertise in the area of imaging techniques is still being held within in library.

## Quality Assurance

Because of the high standards for digitisation, quality assurance of the digital files was automatically on a high level as well. From the beginning quality control managers checked everything that was digitised. Sometimes, in the first small-scale projects, every file was checked and had to be scanned again if the quality was not perfect.

For the first large text digitisation project, the Parliamentary Papers, a quality procedure was set up to check everything: the quality of the images and the OCR, the integrity of the XML files and the way in which al different files were connected to each other. For this purpose a dedicated tool was developed for the project. The costs for development of the tool amounted to over € 200,000 and about € 80,000 per year for maintenance. In this project quality assurance took up about 10 percent of the available budget.

Such a high level of quality assurance is not feasible for all large-scale digitisation projects. It is not realistic to check the quality of all files and even if only a sample is checked, it has to be done as automatically as possible. The Koninklijke Bibliotheek is therefore developing tools to assist in automatic quality assurance.

Initially, a lot of effort was spent on quality assurance of OCR. For every batch of 50,000 scanned pages a sample of one percent was taken and checked character for character by project staff. Unfortunately, when quality standards for OCR were not met, there was little a supplier could do, because at this moment the quality of OCR software for historical documents is just not good enough and manual correction of the text is not feasible. For these reasons quality assurance of OCR has been abandoned for the time being.

### Storage and Long-term Preservation

At the start of digitisation activities at the Koninklijke Bibliotheek the master files were stored on CD-ROMs and DVDs. Thousands of them were loaded into the presentation system and the CDs were saved as a backup. By now everyone will agree that this kind of storage is not a good idea. Examples of corrupt CDs are well known.

A few years ago it was decided to store the master files of digitisation projects in the e-Depot, the digital archiving environment of the Koninklijke Bibliotheek that ensures long-term access to the national electronic deposit collection. However, to store the master files of all digitisation projects about 1 Petabyte of storage capacity is needed and storage in the e-Depot is very expensive. Until recently there was a policy to store all master files of all digitisation projects in the e-Depot. Because of high costs the storage strategy had to be changed so to strike a balance between costs, access and preservation. At the same time research is being done into alternative file formats to minimize the need for storage and the Koninklijke Bibliotheek is considering to use only one file both as master and access file.

### Costs

In earlier projects nobody really knew what the precise costs for the several steps in the digitisation process amounted to. It was therefore difficult to do financial planning for new projects. Now, all costs are made visible, including all kinds of overhead costs. It is now clear which aspects of the workflow absorb most of the budget, making it easier to handle those aspects more efficiently. At the Koninklijke Bibliotheek the actual digitisation now takes up only 50 percent of the available budget. A few years ago this percentage was even lower: only 30 to 40 percent was available for digitisation while a high percentage of the budget was needed for analyses of the material and quality assurance. It is desirable that the available budget for actual digitisation rises up to 70 to 80 percent. As mentioned earlier, exploitation costs are becoming ‘dramatic’ and this problem also needs to be solved. At the

Koninklijke Bibliotheek costs are shared between different projects by joining similar activities. And – although no funding is received from commercial partners – new business models are being explored, such as, for example, digitisation on demand.

### Organisation and Project Management

At the Koninklijke Bibliotheek all digitisation activities take place in the R&D Department while other departments of the library are not really involved. Because digitisation activities gradually become a regular task of the library it is important to integrate these activities with the other departments of the library.

All digitisation activities are project-based. Now that the scale of digitisation projects is growing, each project has become a department of its own, with different project managers for similar activities like selection and digitisation. At the moment seven very large project teams are in place and work has to be rearranged to prevent each project from reinventing the wheel.

## Conclusion

After more than ten years of digitisation activities, a lot has changed, but many libraries are digitising as though it is still 1995. Digitisation seems to be in a transition period in which old assumptions have to be reconsidered. A better balance between quality, quantity and costs has to be struck if libraries wish to digitise on a large-scale. Digitisation processes have to become more efficient and the only way to do this is to limit expectations and not try to be perfect at all costs.

*February 2008*

## Websites Referred To In The Text

Memory of the Netherlands, http://www.memoryofthenetherlands.nl

## Notes

[1] See for example 'One Hundred Hightlights', http://www.kb.nl/galerie/100hoogtepunten/index-en.html

[2] This figure for the number of objects to be digitised in Metamorfoze are based on a rough estimation.

[3] Mass Digitization: Implications for Information Policy. Report from 'Scholarship and Libraries in Transition: A Dialogue about the Impacts of Mass Digitization Projects'. Symposium held on March 10-11, 2006, University of Michigan, Ann Arbor (U.S. National Commission on Libraries and Information Science, May 9, 2006) p. 11.