A practical approach: the quest for a large-scale newspaper digitization workflow

Edwin Klijn, Project Manager, Koninklijke Bibliotheek (National Library of the Netherlands

Abstract

In 2006 the Koninklijke Bibliotheek (National Library of the Netherlands) set up a production chain to digitize 8 million newspaper pages in five years' time. Although the library had gained valuable experience with digitizing the Dutch parliamentary papers, the project Databank of Digital Daily Newspapers (DDD) – the largest digitization project in the history of the KB – took everything to another level: processing an average of 200,000 pages per month put high demands on the whole organization, in terms of staff, IT-infrastructure, technical solutions and workflow management. Digitizing on this scale inevitably means moving from state-of-the-art, 'boutique' digitization towards a workflow where quality is counterbalanced by arguments of quantity. This paper is about the practical choices made in this context by a heritage institution that – after nearly 15 years of experience with digital – is rapidly transforming into a truly digital library.

A short history of digitization at the KB

The library made its first exploration into the digital in the early 1990s. As the internet was gradually appealing to a growing, worldwide audience, the KB launched its first 'serious' website in 1994. The main purpose of this web service was to provide basic information about the library and its holdings. By way of appetizer, in 1995 a web exhibition showed images of highlights from the collection. Library staff – most of them originally involved in microfilming or photography – learnt step by step about how to create digital images. Metadata were often hardcoded into the HTML-pages. Digitization took place on a small scale, focused on the visually attractive objects and was done by self educated 'image scientists' within the library.

The start of the program Memory of the Netherlands in 2001 was a next milestone. This national digitization initiative, supervised by the KB, was set up to digitize and put online Dutch cultural heritage objects from the collections of a wide variety of heritage institutions. On the basis of technical guidelines participating partners in this program were expected to take care of the digitization themselves. The role of the KB – to provide clear, uniform technical specifications – stimulated in-house expertise on digital imaging and metadata. Just like the library's web exhibition the Memory of the Netherlands concentrated on visually attractive images.

In 2003 the KB started a project to digitize the Dutch Parliamentary Papers (1814-1995). With a vast amount of 2.3 million pages it was the library's first mass digitization project. It was destined to fulfill a pioneering role for many similar projects to follow: text oriented and targeted at creating a regular, stable workflow from paper original to website, enabling us to digitize large quantity in a short time. The size of the project forced the library to make pragmatic choices. Metadata needed to be more standardized than ever before. Customized QA of every single file was simply impossible. The experience gained from the first experiments and also from the Memory of the Netherlands was now used to instruct the scanning company to which the digitization work was outsourced.

Putting this huge amount of data online lead to new demands on the library's IT infrastructure. As it was foreseen that similar mass digitization projects would follow, generic services were developed, for instance to highlight a hit term in an image. All files were stored in a permanent, consistent way using unique resource identifiers. In 2002 – in order to keep electronic resources over time – the library built the e-Depot, its own digital preservation system.

In 2007 the KB received funding from the Ministry of Education, Culture and Science to set up a large national newspaper digitization project. The project Databank of Digital Daily newspapers (DDD) will digitize 8 million pages before 2012, from the oldest one (1618) to modern newspapers (until 1995). It has taken mass digitization of text materials to a next level, benefiting from the experiences of the Parliamentary Papers as well as all other previous digitization expertise that has been acquired throughout the years. DDD is setting new standards, not only in terms of production numbers but also in being the first mass digitization project in the Netherlands that has entered negotiations with representative bodies of all right holders to get permission to digitize up until 1995.

The newspaper digitization workflow goes through a process of selection, preparation for digitization, digitization, quality control and presentation. These steps are exemplary for other projects, as well as for a digitization chain in general. Some of these practical choices made are briefly discussed in this paper to illustrate the difficulty – and at the same time challenge – of finding a right balance between quantity and quality.

Selection

The newspapers are nominated for digitization by a Scientific Advisory Committee, consisting of prominent (media)historians, linguists, journalists and representatives of other potential user groups. Currently about 1300 national, local, regional and colonial newspapers have been shortlisted. Some of them are available both on microfilm and paper original. Although according to our own research scans from paper originals produce relatively better OCR, the library still prioritizes scanning from microfilm to other alternatives. The main reason is that the relatively small gain in quality does not fully counterbalance the extra scanning costs when digitizing from original (about four times as expensive). Also, for some microfilms a lot of effort has been put into gathering a complete set of newspapers to be filmed. This time consuming work will have to be redone if one chooses to scan from the paper original.

The budget saved by digitizing from microfilm is reallocated to do more newspapers or for instance to compensate for relatively expensive scans made from 17th and 18th century unique newspapers.

Intellectual Property Rights

Dutch law regulates that authors hold the legal rights of their work until 70 years after their death. Also, publishers own rights until 70 years after publication. In principle nearly all twentieth century newspapers in the selection (an estimated 2.5 million pages) can be affected by claims from legal right holders: journalists, illustrators, photographers, publishers and others. Republishing these newspapers on the internet cannot be done without the explicit permission of all right holders involved. According to Dutch law one is obliged to make a 'considerable' effort to retrieve all rightful stakeholders and receive their consent in advance. In case of mass digitization it is an extremely time consuming, if not practically impossible task.

Currently, the KB is involved in negotiations with the representative bodies of all right holders (publishers, journalists, etc.) to reach a bulk arrangement: the library – in exchange for financial compensation – is being relieved from its legal duty to track down all right holders in advance. The newspaper project is a test case for all other future digitization efforts that the library – and also many other heritage institutions in The Netherlands – will undertake.

Preparation for digitization

The digitization work is outsourced to a commercial party. Before the original newspapers – paper based, microfilm or already digital – are transported to the vendor in Kampen, every individual issue is described in the database. Also, small repairs are performed to make the newspapers ready for scanning. Every week about 50,000 pages are prepared for digitization by a special library department. Biggest issue so far have been newspaper bindings that have been too tightly bound, causing text loss. If there are too many irregularities and there are no alternative copies available these newspapers are excluded from the selection. If only a small portion of the page is affected, text loss is sometimes taken for granted.

Scanning, Optical Character Recognition (OCR) and metadata

The project spends about three thirds of the overall budget for digitization on metadata and OCR. Apart from a full text simple search, users will also be able to perform smart searches by narrowing down on time span, place of publication, distribution area and/or article class (advertisement, family notice, illustration with caption, news item). All headlines are manually corrected to 99.8 word accuracy. The project management chose to invest in manual headline correction because of the expected low quality OCR text from some 17th and 18th century newspapers with old Dutch and Gothic or Fraktur fonts. We choose to having all articles divided up into classes because of the relatively low costs in combination with the huge access improvement for our future users.

The file format for master images is JPEG2000, lossless compressed. Considering the ambitions of the library to do more mass digitization in 2006-7 it became evident that maintaining the policy of storing everything in TIFF would obstruct our plans for the future to digitize large quantities. According to a calculation by the IT-department storage costs for maintaining 1TB were estimated at 8500 for long term preservation (e-Depot) and 7500 for web use per year. The KB did research on alternative file formats for master files. A comparison in terms of storage capacity, image quality, long term sustainability and functionality was made between JPEG2000, PNG, TIFF LZW and JPEG quality 10. JPEG2000 was recommended as a good option in the specific situation of the KB and is currently the library's standard for master files in all digitization projects including DDD.

Other choices made in the project are in line with the library's policy: we prefer to stick to open standards, try to avoid vendor lock-ins and ask files that fit into our IT-infrastructure, to facilitate re-use by third parties.

Quality control

Every week an average of 50,000 pages are digitized. The vendor delivers all files (images, OCR, ALTO, article texts, MPEG21, PDF) to the library on a weekly basis. There are currently two ways of checking: automated and manual. The XML files are automatically checked for their well-formedness and validness. Also the file naming and the structural metadata are both checked by way of scripts. The article segmentation and headline correction are checked manually by sampling. The images – mainly the targets (made for every scanner for every shift on a daily basis) – are checked semi-automated by our quality control managers. If one of the automated or manual checks do not meet the requirements the whole batch is rejected and returned to the supplier.

The QA is monitored by keeping track of the time especially the manual checks take. So far by having the same people doing the same checks repeatedly we require about 6 fulltime staff capacity to process all files for an average of 50,000 pages a week. One of the biggest issues is and has always been the time it takes to copy all files from the hard disks of our supplier to our own system and then – after acceptance – to our web and long term preservation environment. Although copying does not require a lot of manpower it does put a huge load stress on the internal storage system.

Currently our quality control managers are involved in transforming the QA of the images into a fully automated process. This would be another step forward since it would allow less specialized staff members to perform image QA tasks. Another and presumably even bigger advantage is that our vendors can use the same software to check their own equipment at a very early stage. Chances that corrupt data still gets through, are considerably reduced, which will also enable us to do less quality control on the library's side.

Presentation

The newspaper project is the last project that will have its own website. The library is moving towards a more integrated way of presenting its digitized materials,. The basic idea is to create uniform web services for all books, magazines, newspapers and other materials that have been or will be digitized.

A first version of the website was launched on May 27 2010. The functional design of the website has been devised after consultation of the Scientific Advisory Committee, additional user

consultation research amongst potential users and a usability test with some test persons. Our aim was to create a simple, basic website that functions as a solid starting point for adding more advanced functionality in the future. Since the website is expected to attract a large audience precautions have been made to cope with the expected traffic. Components in our IT infrastructure have been added or scaled up. A load balancing configuration was implemented to automatically split the user requests between the available capacity at all time.

Choices we had to make on the access part of the workflow mainly involved relatively basic questions like: Are we going to present the (sometimes poor quality) OCR text to our users? What is the maximum size of the PDFs that we are going to offer? Should they be in color or grayscale? How are we going to highlight the hit terms in the images?

At the very beginning of the project – when the specifications for the European tender for digitization had to be made – many important decisions were taken that had consequences throughout the whole workflow. Apart from the known issues we also encountered a few along the road. To have a smoother workflow at our vendor's side we decided to separate the delivery sequence of the originals from the delivery sequence of the digital files. Now however we only know exactly what has been scanned until the data has been indexed – one of the last steps in the whole process. This makes it difficult to have full control over which title is put online at what time.

Where do we go from here?

The ambition of the KB is to make digitally available another 60 million pages (books, magazines, newspapers) before 2014. The library is considering new strategies like for instance digitization-on-demand, cooperation with commercial parties and an in-house digitization production chain. In many ways DDD has been and will be paving the way for all future digitization activities of the library. One of the most crucial shifts will be to embed an efficient digitization workflow within the organizational structure of the KB. The library already has a department responsible for preparing originals before they are digitized, a cross-project QA-team to check the images and the metadata and an IT-infrastructure that stores and retrieves the data in a generic way. Nevertheless, even more effort will be required in the next few years to standardize and improve the digitization workflow, for example by reducing the costs of required manpower and spending a higher percentage of the budget on cheaper digitization. Whatever the future brings, the credo remains the same as it has always been: practice makes perfect, or at least as perfect as it is practically possible.

Author Biography

Edwin Klijn is Project Manager for the project Databank Digital Daily Newspapers at the Koninklijke Bibliotheek (National Library of the Netherlands). From 1999 he has been involved in several digitization and preservation projects for the European Commission on Preservation and Access (ECPA) and the Royal Netherlands Academy of Arts and Sciences (KNAW). He is co-author of 'In the Picture. Digitization and Preservation of European Photographic Collections' (2000), 'SEPIADES. Recommendations for cataloguing photographic collections' (2003) and 'Tracking the Reel World. A survey of audiovisual collections in Europe' (2008). He also published several articles on newspaper digitization.