Digitizing Simplified: large-scale digitizing of documents at the customers' request

Marc Holtman; Stadsarchief Amsterdam (SAA); the Netherlands

Abstract

The SAA digitizes archival documents on customers request. Delivery time has to be short and costs must be as low as possible. Therefore our standard of quality is based only on user requirements. We developed a working process based on large scale to allow a more efficient work process.

The Archiefbank: a system for consulting and downloading archival documents using the web

All archive inventories of the SAA (Amsterdam City Archives) have been available for consulting on the website since 2004. This search system, thanks in part to the availability of the data to search engines like Google, was being used intensively immediately after its launch. Yet from the point of view of the user this system was limited: the inventories described the content of the archives, but in order to consult the documents themselves it was still necessary to visit the reading room.

For this reason the SAA began to develop the Archiefbank, a system for consulting documents and downloading them from the web in 2006. The first version was launched in July 2007 giving access to 1.500.000 scans. From August 2007 onwards this is augmented with a weekly production of 10.000 scans, primary based on customer requests.

Digitizing at the customer's request

The SAA manages 32 kilometres of archives. Digitalizing every sheet of paper would mean an incredibly large amount of paper. So where do we start? Public favourites seem to be the obvious choice, but research into statistical data from all the requests for originals over the last five years points out that there is no such thing.

The starting point in the Archiefbank is therefore to get the *user* to prioritize the digitizing. In practice this means that all public inventory numbers can be requested for digitizing via the search system. In principle all requests are honoured.

Low cost and short delivery time

Archival research easily runs into dozens or even hundreds of documents. Our customers will need large quantities of scans, so the price of an ordinary Xerox copy has to be the benchmark for the price of a scan. This means the costs of production have to be very low. This obviously has implications for the reproduction process: an entirely new reproduction process had to be developed, with *large scale* and *quality geared to the objective* (that is: consulting digitized documents directly from screen or in print) as its key concepts.

In addition to this the delivery time for digitizing on request will have to remain short. This can only be achieved when two requirements are met:

- a streamlined, efficiently organized, large-scale reproduction process
- a very fast and fool-proof web application

Standard of quality based only on user requirements

In order to keep costs low the standard of quality is to follow from the requirements that the *end user* places on the scans, and *not* to assume more than that. A level of quality higher than necessary for the purpose has no additional value for the customer; all it does, inevitably, is increase both incidental and structural costs, and this makes no sense within the framework of the Archiefbank.

The purpose of the scans produced in this reproduction process is archival research using the web. The standard of quality the scans must meet is that the information legible in the originals must also be legible in the scans when displayed on an ordinary screen or printed on an ordinary printer. Reproduction of details which are not part of the textual information, such as the structure of the paper, is not required. The standards required for colour reproduction is not particularly high. Obviously red has to stay red, but whether the shade exactly represents that of the original is of less importance. It adds no value at all to the reading of the information.

The documents that are being digitized in this reproduction process can have the following forms:

- small and large size (up to the size of an open newspaper)
- bound and loose-leafed entities
- card indexes
- old and modern material
- low and high contrast documents
- text alone, text and image together
- hybrid forms

The same standard of quality applies to all documents, so when an inventory number contains images the 'legibility of the visual information' will still be the norm. This is obviously a somewhat difficult concept to interpret with, for example, photographs but the basic principle is that in every case a complete reproduction of the subject of the photograph ought to be possible. When details in the background are lost as a consequence of the relatively low standard of quality for images this does not really matter. On the face of it this seems to be a somewhat strange assumption ('surely you want to see everything?'), but it does actually follow from the use of the original in archive research (and *that* is the purpose of the exercise). Firstly the album is skimmed through to get an impression of the content, and then a number of photographs may be selected for reproduction in high quality and scans made for publication or framing. The costs of these reproductions are therefore considerably higher than for an ordinary photocopy from the album.

Obviously skimping on the quality of scans of images (it can be better) is purely an economic decision, not one taken on principle. At present the costs of digitizing to photograph quality are on average ten times higher than scanning to the standard of legibility. If these costs were to be passed on the customer, digitizing a photograph album, for example, would become so expensive that it would be beyond the reach of many people. The experience of the SAA is that in the reading room the average user opts for an ordinary copy, possibly in colour, and only occasionally people place made-to-measure orders for high quality reproductions.

If digitizing is being done with another objective, for example scanning for inclusion in the Beeldbank application (which contains high-resolution images of photographs, drawings en prints) and providing high-quality reproductions as a direct download, different requirements will have to be set and an entirely different reproduction process may have to be used. The form in which the scans are supplied will generally be different too. In an image databank, for example, you can choose a selection of a number of photographs from an album; these can be cropped as individual scans, and described separately. When this is digitized for the Archiefbank the complete album will always be done, including the front and back of the album; the quality will be average, but the customer will be able to afford it.

From this it can also be concluded that the distinction between a request for images in the Beeldbank, where the customer usually only buys just a few scans but demands high quality and the purpose is most often publication or exhibition, and a request for archive material (Archiefbank), where legibility is what counts and where sometimes dozens if not hundreds of scans are being purchased at a time, is without a doubt logical and justified. But this can mean that the original document qualifies for both systems. At the SAA it works like this: if originals are being digitized for the Beeldbank on the basis of our own selection process, the standards for images apply. If they are being digitized at the request of the user, the Archiefbank standards apply, the argument for this being the manageability of costs for the customer. If an inventory number digitized for the Archiefbank is to be recorded later in the Beeldbank it is digitized again.

No storage of separate, uncompressed (or lossless compressed) images

The scans are made for the purposes of archive research by the user and *not* as a substitute for the originals. Conservation of the *original* remains the major concern. Digitizing archive items for the purpose of use however, does have a real conservation function because the originals, in principle, no longer have to leave the repository.

The SAA makes no distinction in quality between master- and accessfiles in the storage of the scans. For the time being the SAA is opting for storage in JPEG format, with a compression factor of 10 (Photoshop). This results in an image with:

- a relatively small filesize
- first-rate reproduction in detail and colour, and hence
- an extremely legible image on screen and in print, and
- enough flexibility for manipulation for the purposes of functionality in the Archiefbank document viewer

Permanency of the files is guaranteed by redundancy in storage. For the moment the files are saved internally and on the web server. Internally only the masterfiles are stored. When the images are uploaded into the web application automatic derivatives are made for the purpose of:

- zooming in on the various layers
- allowing the individual user to determine the degree of contrast

These derivatives are only stored on the web server. The image that the user is able to download has the same quality as the master image supplied by the digitizer. In due course it can be expected that these standards will be adapted with the introduction of new storage and compression formats.

The reproduction process

In a pilot project it quite quickly became obvious that the available guidelines and best practices for digitizing often take no account of the enormous number of scans to be produced when digitizing documents. The SAA, in close cooperation with a digitizing partner, has developed and implemented in the organisation a reproduction process that can manage an average of 10.000 scans a week varying from complicated to simple documents, all in a standard quality and at an absolute bottom price.

The main principles in this reproduction process are:

1. Identification of the units to be digitized is based on order numbers

A unit to be digitized must be able to be identified at each step of the handling process. The existing and physically present combination of an access number / inventory number should in theory be able to be used for this, but because of inconsistencies in the two numbers this does not provide a truly consistent (and hence clear) label. As a consequence of differently applied forms of labelling and the use of different types of labels the identification of the originals based on these two existing numbers can be difficult in practice.

In any event there should be a simple way for the digitizer to be clear what he is dealing with, and no doubt should exist about the units to be digitized.
The SAA therefore decided to give units to be digitized a unique, meaningless order number. This number is used throughout the entire process and forms the basis for:

- communication with the digitizer
- scanning
- attaching file names
- computerized registration of the file names in the management systems
- billing

Order numbers are issued automatically through the internal SAA management systems.

2. File names of the scans are meaningless

The basic principle at the SAA is that file names may not contain any meaning that signifies content. For example, file names do not include an inventory number, archive number or other types of information that could be subject to change in practice. A change in this information would in fact mean that the file names of *all* images concerned would have to be changed. This also goes for image files stored offline on DVD.

Prior to scanning—in contrast to what is often the standard when digitizing images—the originals are not given the file name item by item, for instance by way of a barcode. Nonetheless when scanning it has to be obvious which file names should be attached, and it has to be possible to register this automatically to the relevant inventory number in the management systems.

The most obvious option is to register file names before digitizing, but this has a practical drawback because it is never certain in advance how many scans there are in an inventory number (the number of sheets is not necessarily the same as the number of *scans*). The SAA has opted for a solution where the order number is part of the file name. A file name is always made up of twelve characters, the first six being reserved for the order number. The next six characters contain a serial number starting with 000001. Prior to scanning, an order ticket is added to each of the inventory numbers, on which, among other things, the first file name is given, for example A02043000001. After scanning the first sheet the number is increased. Strictly speaking a range of 999999 possible file names can be attached to each order. This is more than adequate, and in practice even excessive.

The significance of the order number in the file name disappears at the moment that the file names are registered in the management system. Therefore it is a matter of meaningless file names, so at a later point in time the images can be attached to other inventory numbers without any problems, or spread over a number of inventory numbers. The sequence can also be changed without having to amend the file names.

3. Always a complete inventory number is scanned

By definition the entire inventory number is digitized, never just a selection of pages. There a few simple reasons for this:

- The costs of scanning are not so much a factor of the quantity, but rather of the manual processing involved in it. Therefore it is most advantageous to scan an inventory number in its entirety at one time.
- If only a part is being digitized, it has to be indicated in the originals or in the metadata (which can only be done if the originals have been numbered continuously, and this is usually not the case) which part should, and which part should not be digitized. This is a time-consuming task in the preparation, and makes the scanning process particularly complex, as it must be clear without any doubt what should and what should not be scanned.
- When scans are shown in the web application behind an inventory number the customer expects everything to be there. It is entirely illogical if this is only a part of the total.
- The entire preparatory process has to be gone through once again when the non-scanned pages have to be digitized later.

4. Material preparation as simple as possible

An inventory number is digitized as it is found in the repository. Only a rough check of material and content takes place there. Likewise the sequence is not checked in detail. The instruction to the digitizer is to keep the sequence of the originals when scanning. Should it turn out later that the sequence is wrong, then if need be the *scans* can still be put into the correct order.

5. The originals are not numbered

Numbering the originals prior to digitizing has the advantage that the completeness of the scans compared to the originals and the completeness of the originals (at the time of numbering) can be guaranteed. But this is only true if the numbering tallies *exactly*. An error in the numbering is far worse than not numbering at all. If the next number is not in order, because, for example a number has been accidentally missed out (… 9, 10, 12, 13 …), this will *always* create problems.

The added value of numbering compared to not numbering is slight. The original file can always be referred to if there is any doubt about the completeness of the scans. Furthermore digitizing leads to a high level of certainty that the original file is complete and remains so: the originals no longer have to be used.

6. No fragile material or material packaged in an unconventional way

Material—like charters—that is very fragile and therefore requires special handling when packaging, transporting and scanning falls outside the reproduction process described in this report. Items like these can therefore not be digitized on request. The scans will of course be offered through the Archiefbank if they are going to be digitized in a separate project (for example the oldest charters from the Iron Chapel in the Oude Kerk in Amsterdam were digitized in a project like this and then entered into the Archiefbank).

7. Constant production

A large-scale reproduction process can only be organized effectively when constant production is assumed. It is then clear to the archive institution how many hours work a week is involved in preparation and finishing, and the digitizer can plan the best use of scanner tables and personnel.

The basic principle of the SAA is to work with fixed weekly batches of 10,000 scans. If too few scans are being made from customer requests this is complemented with our own selections. When there are too many scans, some of the requests will be sent with the next batch (the delivery time to the customer is of course longer).

8. Contracting out the scanning to an outside partner

It is true that the SAA has professional facilities in-house for analogue and digital photography, but these are not designed for large-scale digitizing of archive material. The complexity of the type of material to be digitized calls for specialized scan set-ups, hardware, software, knowledge and technical infrastructure. Investing in this only makes sense where there is very high production in a commercial set-up and it can be organized on a large scale. Contracting out of this process was therefore a logical choice for the SAA.

There are a great many scanning companies with a wide variation in charges, scales of fees and approaches. It was established that many scanning companies do have experience in bulk processing but not of the degree of complexity and diversity involved in large-scale scanning of the sort of archive material managed by archive services.

In order to get the scanning of large amounts of material done quickly, the processes of the archive service and digitizer will have to dovetail, and the operations will have to be integrated into the process at a logical moment. For instance, checking of completeness (scans compared to originals) is an essential step in the reproduction process. The most logical thing is to get the digitizer to do this immediately after scanning and not to wait until the originals come back to the archive service, as only then can remedial action be taken quickly and efficiently on discovery of missing scans. However this calls for a high degree of trust, and means that contracting out the scanning is more than just awarding a contract to a supplier.

9. What is digitized is available to all users

All digitized inventory numbers are available to *all* users through the web application. This also applies to inventory numbers that are digitized at the request of the user. So digitizing on request is not done exclusively for the person who requests it. Communication about a request can therefore be kept to a minimum. Making the request is enough, there is no quotation made that would only slow down the process. If it turns out that the person who made the request does not want to buy it after digitizing this is not a problem because the scans are immediately available for other users anyway (the person who made the request does not even have to be the first buyer).

A web application with user-friendly front- and back office

An efficient procedure of large scale digitization on customer's request starts and ends in the search-system of the archive inventories on the website.

FRONT OFFICE

On the internet, service has to be fast because a web user is impatient. He wants to find a few scans out of millions of scans and he wants to find them in less than no time. Applications for digitizing and buying scans should be provided in a simple and rapid method.

Experience with delivering scans to customers in a pilot project previous to the construction of the Archiefbank has shown that legibility is only relative. The average user is not at all interested in a perfect copy of the original: they want the scan to be as legible as possible. In general, A high contrast image is considered as easy-to-read but this surely does not apply to everyone. Those personal preferences are met by means of a viewer, designed especially for the reading of documents. A master JPEG scan is saved in the web application in duplicate:

- The master image
- A derived image on which a curve has been applied, resulting in an extremely high contrast image.

Laid exactly on top of each other, both scans will be shown in the document viewer. A control slider allows the user to adjust the transparency of both images, resulting in the arrangement of contrast as desired.

The JPEG images in the application are also tiled and saved/ stored in different resolutions. The ability to zoom in on the images is based on the different layers of resolution.

The document viewer is based on HTML. Plug-ins or additional software is not necessary.

BACK-OFFICE

In the back- office we want immediate and safe access to all management-information on scans, import-routine, users and sales. These tasks can be executed from the CMS of the web application.

The work is further supported by an internally developed workflow system in which all relevant information is readily available and reports are generated.

Conclusion

Large scale digitalization of archives at lower costs and quicker delivery is possible if:

- The image quality is based on user requirements and no more than that,
- The reproduction process is streamlined, efficiently organized and automated wherever possible,
- The back-and front-office systems are user- friendly and fit the demands and needs the customer and employee put on it optimally.

References

[1] M. Holtman, Digitizing simplified; Large-scale digitizing for archive research. (Augus 2007). Available at: http://stadsarchief.amsterdam.nl/stadsarchief/over_ons/projecten_en_jaarverslagen/

Author Biography

Marc Holtman is senior digital public services at the SAA. He started working there in 2001 as project leader for the realisation of the Beeldbank *(image bank). After that he became project leader for of the online inventories and Archiefbank (application and reproduction process). At this moment he is the coordinator of all digitalization projects within the SAA.*