ARCHIVING CONFERENCE
IS&T imaging.org
https://www.ingentaconnect.com/contentone/ist/ac/2006/00002006/00000001/art00051

Achieving Quality in Digitisation Workflow

Rob Mildren; National Archives of Scotland, UK

Abstract

In 2000 the National Archives of Scotland embarked on a project to digitise manuscript material dating from 1500 to 1901. This project was undertaken in conjunction with the Genealogical Society of Utah (GSU). The requirements of this project led us to implement techniques and procedures that have allowed us to capture more than 4 million images. This paper describes the various aspects of the workflow and in particular how we considered quality issues at each step..

Selection and assessment of material

It was important that the digitisation project paid due regard to preservation concerns. Conservation staff were employed exclusively for the project, and had a major role to play in advising on the selection of equipment to be used. They assessed the physical state of the volumes and warrants (original wills and inventories in loose leaf format) in advance of digitisation. This process was undertaken sufficiently far in advance to ensure an adequate body of material for the digitisation to proceed without delays. They carried out conservation on documents where necessary and employed the principle that intervention would only be required where either the image would be significantly enhanced - for example if the pages were very dirty - or where without conservation input, the digitisation process could cause damage to the manuscript[1].

Once the digital images were created the conservation staff created custom boxes for the proper housing of the volumes and these were then placed in good storage conditions.

Preparation and pagination

There were additional staff resources for conservation of the material (before and after digitisation) and several approaches to loose leaf and bound material were developed.

An important part of the preparation process was ensuring that each page to be digitised had an accurate page number. This was then incorporated into the document reference to form the file name. Conservation staff paginated all the early material up to 1750, but the later material was paginated by our team of volunteer camera operators according to guidelines laid down by the project archivists and conservators. The pagination process helped to define the file name for the digital image but it was also an important indicator that the camera operators used to ensure that

- all pages were captured
- no pages were duplicated
- no images were missed

The accuracy of the page number was one of the key checks carried out by the quality control operators.

Proper handling by trained camera operators

The conservation staff also established handling guidelines that all the camera operators were required to follow and also gave operators training in handling the documents to minimise damage and to recognise where further conservation input might be required. The requirements to undertake training and abide by the handling guidelines were an important part of the contractual relationship with the GSU.

Image capture software that minimised operator intervention

We needed to develop a system which would allow staff, many of whom had little or no ICT experience, to concentrate on their task of capturing accurate, good quality images and to do so at a good rate of throughput. We therefore looked to simplify the steps involved so that, once a volume had been set up on the book cradle to be digitised and the metadata for the volume entered, the camera operator had a one button approach to capture each of the images that followed. This therefore included automatically naming the file and storing them away. Images were cropped automatically, if required, and checks were made on the colour to highlight anomalies.

Image capture itself was quick. We used a greyscale camera and attached a computer controlled filter to it. The camera took three pictures with the red, green and blue filters and then combined them to display a composite colour image on screen for the operator to check. Once the final image had been captured the system would start to save the image and also released the book cradle to allow the operator to turn the next page. Each image would take about 3.5 seconds so a full colour image, with three takes, would take around 11 seconds. Allowing for the operator to check the image and turn the page this means a full cycle time of around 15 seconds per colour image.

Images were captured as colour tiff images onto the hard disk of the local PC. This minimised any network traffic and meant we could invest in fast disks with a large capacity for each of the six camera PCs we had purchased. As we saved the images only in TIFF format we had no overhead at the pint of capture for the creation of any other file formats. This operation took place once the camera operators had completed their work for the day and we would run the image format program overnight. In order to manage the large number of images produced we kept to a naming

convention based on the original file reference plus the page number. This file reference was also used to create a directory on the server to store all the images for a particular volume and meant that it was straightforward to name and find an image for any page for any volume.

Image Quality – Fit for Purpose

We used digital cameras rather than scanners for the digital capture. Digital cameras operate by focusing the image on a light sensitive chip called a CCD (Charged Couple Device). The CCD has a fixed capacity and for the two cameras we operated for this project the arrays were the following sizes

Camera type	CCD Size	Total Available Pixels
Kodak Megaplus 6.3i	3072 x 2048	6291456
Atmel Camelia	3500 x 2300	8050000

So regardless of the size of the document being digitised we are limited by this capacity. Line scanners operate differently and move a line array CCD across the document to a fixed size. The optical resolution is therefore normally expressed as dots per inch (dpi). With a fixed CCD capacity then the resolution would be different depending on the size of the document being digitised. To achieve an equivalent resolution of 300 dpi would mean restricting documents to less than 10 inches by 7 inches. In order to meet our requirement the image quality had to be "fit for purpose". Our purpose was to make the documents legible on screen or on printout.

We needed a different metric that demonstrated sufficient quality but was suitable to the various sizes of documents we had to digitise. We agreed on a standard whereby the pen strokes of the handwriting were examined. The number of distinct pixels for different types of line thicknesses was measured and we concluded that if we had 3-4 pixels for each normal line then, regardless of the use of image on screen or on a printout, that we had captured sufficient information to represent the image accurately. This conclusion meant that we could capture images of an open volume rather than having to take images of each page on either side of the volume. This obviously increased the throughput but also halved the strain on the documents that would have been required if we had taken each page individually.

The images were tested by our user group and found to be very acceptable and judged to be of a high quality and sufficient for their needs.

Formal quality control procedures

Quality control was undertaken in a separate programme. Once images had been converted to jpeg format, which happened overnight to minimise capture times, quality control was carried out by another operator. Once a volume had been checked the results were recorded. This means that we can ascertain whether an image was examined (and by which operator) or whether it was approved as part of a larger batch. Once complete the quality control program produces a summary printout together with a list of images rejected and the reason for the rejection. We started the project with a 100% check of every image but the most effective results obtained from this program were found to be from a 30% random selection of images per volume.

Part of the function of the quality control process was to highlight and, where possible eliminate, common errors. This led us to improve our procedures for focusing on the document, for calibrating white balance and for holding documents securely whilst the image wa being captured.

Software for data back ups

We retained copies of the colour tiff images on the hard drive of the machine that produced them until the quality control program was complete and any necessary retakes were completed. Once this was done we had simple procedures in place to let operators identify material that had been completed, how much space they would take up on the tape and then write them to tape and also record the information about the tape and starting block on a database.

More recently we have started to use USB hard drives instead of DLT. Over the years the tape drives were a common point of failure in a camera set up.

Resilience

On site image storage is both online and on tape. The online storage (approximately 2.5 terabytes) makes all the jpeg images available. The online storage is protected by RAID 5 and also has a hot spare to immediately fix any disk problems. A full copy of the online storage is held in another NAS building with an overnight copy taking place. A full integrity check ensuring that the content s of the two devices are fully synchronised is carried out each night. Tape storage exists for both uncompressed (TIFF) and compressed (JPEG) colour images. Additional resilience comes from having uncompressed greyscale images written to tape and stored "off-continent" in Salt Lake City.

Procedures for creating links between the finding aid and the images

Once images have been created we needed to provide access to them. A volume of images could include over a thousand pages so giving access to a whole volume would be little help. We didn't have a comprehensive index to all the testaments so had to create one form all the different sources that were available. This included the digital transcription of some published indexes, transcription of index pages from some individual volumes and the creation of indexes where none previously existed. This gives a

direct link between the index and the images referred to in the index. This can only be achieved successfully by accurate pagination of the original document corresponding exactly with the image numbers. Provision has to be made for linking index entries where there is more than one testament per page. This is more common in the pre-18th century registers.

As all the pages were viewed for indexing information this was a second chance to examine images for any faults. The indexing software allowed us to highlight any such images for retakes.

Additional Metadata for the Images

We maintain a database that contains a record for each of the images we have created. This does not include the indexing information used to identify the content of the record but describes information about the processes used to create the image.

During the image capture process we automatically create an entry in a logfile for each image. Attributes captured at this time include the camera id, particular camera settings, date and time of capture, operator id, volume description and indicators whether the image is of a blank page or whether it is a retake. When we process the images to create our derivatives we record the information against the image in our image logfile database. When the images are quality controlled we record the QC information relevant to each image. Once the images are written to tape we record information about the backup device they are stored upon. Taken altogether this gives us a full picture of the creation of the image and is also a key management information tool. The great advantage of our system is that this wealth of information is collected with very little intervention from the operators and causes very little overhead to collect. This information can be exported to a simple text file for single images or for full volumes.

Website for access to the index and images including e-commerce

While we were still capturing the images and linking them to the index, we had planned our e-commerce site to provide remote access. The index would be accessible free of charge, along with a whole range of other supporting information. After undertaking a marketing evaluation we decided that a fixed fee would be suitable, regardless of the number of individual pages that a testament covered. After payment the customer can view or download all of the images relating to a testament and we will retain information about the customer order to allow them to come back to our site and view again the images they had purchased.

This site (www.ScottishDocuments.com) proved a very effective means for promoting access to the images. Since June 2005 the images have been incorporated into the www.ScotlandsPeople.gov.uk website and the original website was suspended until we are ready to launch e-commerce access to the Kirk Session records.

Conclusion

The digitisation process described above has proved very successful at digitising large quantities of original manuscript material in bound volumes. This has led us to undertake even larger projects and we are currently digitising an estimated 8 million pages of Church records. In addition we are considering the modification necessary to allow us to use the same processes to digitise documents on demand.

References

[1] http://www.scan.org.uk/aboutus/Reports/conservationreport..

Author Biography

Rob Mildren is head of ICT at the National Archives of Scotland and was Project Manger for the Scottish Archive Network He was a member of the International Council on Archives Committee on IT with particular interest in Digital Imaging Technology.