Digital Culture & Heritage
Patrimoine & Culture Numérique

Haus der Kulturen der Welt, BERLIN
Aug. 31st - Sept. 2nd, 2004
31 Août - 2 septembre 2004

MANAGING QUANTITY AND QUALITY - DIGITISATION AT THE NATIONAL ARCHIVES OF SCOTLAND

Rob Mildren, Head of ICT, National Archives of Scotland
Hazel Anderson, Team Leader, Scottish Archive Network

http://www.ScottishDocuments.com

Published with the sponsorship of the French Ministry of Culture and Communication

Actes publiés avec le soutien de la Mission de la Recherche et de la Technologie du Ministère de la Culture et de la Communication, France

Interprétation simultanée du colloque et traduction des actes réalisées avec le soutien de l'Agence Intergouvernementale de la Francophonie

Abstract (EN)

The Scottish Archive Network project, supported by the National Archives of Scotland, has completed an ambitious programme of creating digital images of the wills and testaments of Scotland from 1500 to 1901. This project created colour images from the original manuscripts and linked the images to the index entries for each of the people who had registered such a document. This amounted to some 500,000 entries and 3 million pages of images. The digitization project was undertaken in-house in co-operation with volunteers from the Genealogical Society of Utah (GSU) and was completed in around 24 months. This paper describes the main steps involved in ensuring a proper balance between the huge quantity of images to be produced and the need for a high level of quality and demonstrates that large scale digital imaging projects are manageable and can be conducted with due regard to preservation concerns.

Keywords: National Archives of Scotland, Scottish Archive Network, Testaments, Digitization, Genealogical Society of Utah (GSU), digital images.

Zusammenfassung (DE

Das schottische Archive Network (SCAN) hat, mit der Unterstützung des Nationalarchivs von Schottland (NAS), ein ambitioniertes Projekt fertig gestellt, im Rahmen dessen alle schottischen Testamente von 1500 bis 1901 digitalisiert wurden. Es wurden digitale Farbbilder der Originaldokumente gemacht und mit den Indexeinträgen der Personen, die ein solches Dokument registriert hatten, verlinkt. Das Ganze umfasst ca. 500.000 Eintragungen und 3 Million Seiten mit Bildern. Das Digitalisierungsprojekt wurde intern, in Zusammenarbeit mit Freiwilligen der genealogischen Gesellschaft von Utah (GSU), durchgeführt und innerhalb von rund 24 Monaten fertig gestellt. Dieser Beitrag beschreibt die Hauptschritte, die um der Forderung nach einer Balance zwischen der riesigen Anzahl an Bildern und einem hohen Qualitätsniveau gerecht zu werden, bedacht werden müssen. Darüber hinaus soll demonstriert werden, dass groß angelegte Digitalisierungsprojekte zu bewältigen sind und mit gebührender Beachtung von Konservierungsbelangen durchgeführt werden können.

Schlüsselwörter (DE): National Archives of Scotland (NAS), The Scottish Archive Network (SCAN), Testamente, Digitalisierung, genealogische Gesellschaft von Utah (GSU),

digitalisierte Bilder.

Résumé (FR)

Le réseau SCAN (réseau d'archives écossaises), soutenu par les Archives Nationales d'Écosse (NAS), a accompli l'ambitieux projet de numériser tous les testaments des Écossais du 16ème au 19ème siècle (environ 3 millions de pages). Ce programme a produit environ 2 millions d'images numérisées en couleurs d'après les manuscrits originaux, et aussi environ 500 000 entrées d'index. Les images ont été liées à ces entrées. Le projet de numérisation a été entrepris sur les sites de conservation, en coopération avec des volontaires de la Société Généalogique de l'Utah (GSU), et a été accompli en 24 mois. Cette communication décrit les considérations principales pour assurer un équilibre approprié entre la création d'une quantité énorme d'images numérisées et la nécessité de maintenir un niveau de qualité élevé. Elle démontre aussi que les projets de numérisation en masse sont possibles à réaliser sans compromettre la préservation des documents.

Mots clés:
Archives Nationales d'Écosse (NAS), Réseau d'archives écossaises (SCAN), Testaments, Numérisation massive, Société Généalogique de l'Utah (GSU), Numérisation de manuscrits, Bases d'images.

I. Why Digitise?

1. Wills and Testaments in Scotland

Project Background

The concept was to create a resource which would be of maximum value to potential users. Since statistics of interests declared by archive users at the National Archives of Scotland indicated that around half of them were primarily interested in genealogy, the initial solution was a comprehensive index to the wills. Despite the popularity of this group of these records, and their national significance, for they cover the whole of the country, no comprehensive index exists prior to 1876.

However, to provide an improved index to an already very popular range of records would have removed a natural 'bottleneck' on usage and placed an intolerable strain both on the staff of the NAS and on the records themselves, which show signs of wear from overuse in the past few decades. Therefore we concluded that the most effective solution to allow access would be to create digital images, to reduce repetitive demands on staff time, and to ease the pressure on the original documents.

Understanding the Wills and Testaments

The wills, or testaments to give them their proper name, are a unique resource for historical study. Scots law distinguishes between heritable and moveable rights, the former covering land and buildings, the latter covering goods, money and other possessions. Testaments are exclusively concerned with moveable right and consist of three elements. First is the confirmation of the identity of the deceased's moveable estate. If the deceased has nominated an executor, this is known as a testament testamentar; if the deceased died intestate, the executor is appointed by the court and this is known as a testament dative. Second is an inventory of the estate. Third is a will, if the deceased left one.

All three elements of the testament can be of historical and genealogical interest. Normally the executor will be a relation of the deceased, and his or her identity will tell us something about family relationships. In some cases, where the deceased left debts, a creditor could apply to be appointed the executor. If the deceased left a will then there is considerable genealogical interest in what this contains. But perhaps the richest potential lies in the inventories, which in some cases contain full details of the effects of the deceased, with monetary equivalents. They can give us a snapshot of the contents of a house, or a stock of a

merchant, or the tools of a craftsman, together with their values. A wide range of occupations is represented among the recorded testaments, and these are a key source for information on economic and social conditions.

Testaments are not simply the preserve of the wealthy; some are recorded for people of very modest means. It is clear that the testaments exist only for a minority and that in the majority of cases families simply arranged succession themselves, without the bother and expense of the law. For the period covered by the digitised wills (1500-1901) there are 580,000 recorded testaments. These entries fill more than 3,000 volumes and an estimated 3 million pages. The challenge was to produce digital images for all of the documents and link them to a comprehensive index. These index entries and images would be made available on an e-commerce site for purchase over the internet or could be viewed free of charge within the NAS.

II. Managing Quantity

1. Type of Material to be Digitised

Most of the material we captured was in the form of bound volumes. Some material was loose leaf but the bound material represents the overwhelming majority of the material to be captured.

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.

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.

Further information about the conservation input to the project together with recommendations can be found in a published report (http://www.scan.org.uk/aboutus/Reports/conservationreport.pdf).

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 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 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 lace 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	3072x2048	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 4 pixels for each 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. 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.

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.

Resilience

On site image storage is both online and on tape. The online storage (approximately 1.5 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. Tape storage includes 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 may 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.

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) has proved a very effective means for promoting access to the images. 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.