Digitizing and Preserving Records

Richard and Evan Ivie
GeneSys Foundation
Nauvoo, Illinois

INTROCUTION:

Abstract:

In this paper we report on what we have learned from our experience in digitizing several records collections.

Why Go Digital?

The advantages of digital document collections over paper collections are generally well recognized and accepted today. Digital records are much easier to preserve, backup, organize, share and enjoy. And once digitized, the record image will no longer fade and deteriorate. All or part of a collection can be placed on the Internet for access anytime and anywhere. The cost of duplicating and backing up the collection in multiple locations becomes a fraction of the cost of backing up a large paper collection. Digital records provide easier access points and significantly reduced amounts of physical storage space. Digital images also provide accessibility to images by those who can use the documents. This means that family histories will be better documented, with citations and links to the actual images. This brings a new dimension to family history, where not only the conclusions, but also the actual images used in making those conclusions are part of the family historian's record.The biggest plus is however, that organization and searching of a collection is greatly facilitated.

Experience to Date:

Over the past two years we have scanned several records collections. The Nauvoo Restoration Incorporated (NRI) records occupied about 16 4-drawer filing cabinets, 4 large wall cabinets, and perhaps 25 legal boxes of papers. Documents, books, photographs, slides, blueprints, maps, and other memorabilia were digitized. There were about 80,000 images generated. We had three scanner stations and three copy stands for cameras. A large 4 by 4 foot vacuum table was also built to handle large blueprints, maps and other images.

In a separate effort one of the authors scanned his personal and genealogical collection consisting of 3 filing cabinets and 10 boxes of documents. There were about forty thousand genealogical images. He also scanned his financial records (bank statements, receipts, invoices, Internet orders, checks, etc) which consisted of about ten thousand images. When the IRS requested an audit of his records he sent them a DVD which was accepted as a suitable submission. We have also experimented with a number of smaller collections.

The Technology:

Cameras and Scanners

There are two basic technologies for digitizing records: scanners and digital cameras. Scanners provide better control of the ambient light and provide high quality. Digital cameras are more portable and flexible, can handle larger images and other images that sheet feeders cannot handle. They also generally have a faster digitizing rate approaching 1 image per second. This rate can also be achieved by scanners with an automatic document feeder (ADF) if the papers are in good shape and uniform in size.

CCD cameras are sensitive to more than just visible light. With the addition of Infra-red LEDs as a light source different information can be pulled from the records. Digital cameras range from the very portable cell phone all the way to higher end cameras that preserve higher pixels and record the image faster.

Cell phones can use the phone system to transfer the image quickly to remote servers. Digital cameras equipped with Wi-Fi type Compact flash cards can instantly send the image to the world-wide web as well as a laptop's larger screen. The instant feedback to the photographer of the images is essential to preserving good record quality. Looking through large numbers of images that were taken with standard photography shows that just because there was a standardized setup and lighting source does not mean that the picture was not blurry, dark, light or otherwise unreadable.

We used copy stands for the cameras so that we could have better control over the lighting, the focus, and the positions of the documents.

What We Have Learned:

Choice of Scanners versus Cameras:

The scanners we had (three Microtek ArtixScanDI 2010 scanners and a battey of other miscellaneous scanners) had attachments for scanning slides and we used them for that purpose. We had a large collection of maps, blueprints, and other documents that were too large for the scanners and we used cameras for that purpose. The most common camera used was the Canon EOS but we experimented with a variety of others also. Printed, typewritten, and hand-written pages were generally done on the scanners because they created a cleaner background but many were captured by camera at reasonable quality. We did not have a scanner that handled the gutter problem well so many of the books were done by camera. Photographs were done on both the scanner and by camera.

File Format:

We have a permanent archive of all of our images in TIF format. However, for transfer and web publishing we convert the images to JPG or PDF format. The TIF images range in size from 1 MB to 40 MB. The JPG images are in the 10-100K range. The scan resolution used ranges from 250 bps to 2400 bps. Simple printed and type-written documents are generally done at 250-400 bps.

We have generally organized the images into the same structure electronically that they exist in physically: cabinet – drawer – folder – subfolder … About half way through the NRI collection we converted from using separate single page images to using the multi-page image format available with TIF files.

Grouping of Documents:

Each single page or multi-page image was given a descriptive name at the time of digitizing. We have delayed a more exhaustive data tagging (indexing) effort in the interest of capturing the collection digitally.

We broke larger collections of items down into “batches” of about 20-50 items (pages) each. This allowed us to more easily keep track of the flow of pages through the system. We tried to group the pages into logically connected batches so they could be labeled with meaningful names. Where this was not possible we just labeled the collection with a single name and numbered the batches sequentially.

Backup:

We have a three-part recommendation for backing up your scanned files.

- Keep an ***on-line*** copy of your files on one of the commercial backup sites that are available (e.g. www.mozy.com).
- Keep an ***off-site*** copy of the data. We do this by periodically sending copies of the data to interested family members. For smaller collections we use DVD's. For larger collections we use TB external drives for this sharing and distribution.
- Keep the original ***paper*** copies (especially the primary source documents). This may seem contradictory to our goal of going paperless, but until the whole process becomes more stable and trustworthy, it just seems like a good idea …

Future Plans

Image Labeling:

The latest improvements that have been made in OCR and voice recognition technology have been made with statistical engines. The more information is had about the digital data the better the ability to accurately transform the information into a fully digitized electronically searchable form. We hope to provide short demo's on how these technologies aid in tagging and indexing process.

FamilySearch indexing has come up with a very powerful way of indexing records that could be applied to other records as well. The concept of sharing part of the record and using standardized signal sets (these are the names allowed, these are the place names allowed etc.) has made the task of indexing faster and better.

On sensitive records where flash is not allowed longer time-exposure with a larger lens camera significantly helps in the preservation of the record.

Often paper tigers (unsorted documents of varying value) are held by older generations that often have posterity or interested neighbors and younger friends. If these records could be digitized and moved to where they could be indexed by this army of related computer users this could make the records at once better preserved (because they have been shared more locations) and better indexed.

Document Indexing and Searching:

Both Microsoft and Google have advanced picture searching technology. These advances could also be applied to the digital record indexing challenge electronically sorting the pictures.

Microsoft surface (http://www.microsoft.com/SURFACE/index.html) allows a way to quickly work with images, resizing, sorting, and indexing them. The concept of a large projected screen that allows many different images to be viewed simultaneously and sorted all together will allow greater sorting speed. I was able to compare sorting speed from a laptop screen with that of a larger desktop screen and the larger size was definitely easier to sort images as more thumbnails could be seen at once. The 10 finger intuitive touch approach also increases speed of image sorting and hence indexing.

Using voice recognition software allows someone to describe the image as the pictures are being taken, hence the burden of indexing can be done at the same time the pictures are being taken. With statistical information about the images this voice recognition accuracy could be greatly improved. Digital recording could be used to collect the information on the digitized record from living sources.

Scanning Services:

For those who do not have time or a desire to do their own scanning, there are commercial services available. See, for example, www.pixily.com. You can send this company as few as 50 sheets and they will place the scanned images on a passworded site for you and return the documents. Using such a service may actually get the scanning done instead of waiting "until I retire" which never seems to come. There are three basic drawbacks to this service:

- Pixily currently does not handle oversize documents.
- Delicate documents may be damage in transit or during scanning..
- The cost of doing it is 30 cents a sheet which may seem high since scanners are down to $100-200 now.
- I place a high personal value on some of my documents and don't want to trust them out of my sight and possession.

A couple of examples of other possible services include www.acentra.com , and www.neatco.com.

Further Information:

Sites such as http://www.sciencedaily.com/releases/2008/10/081009072208.htm report on new approaches to training the computer to recognize the images and help with the sorting and indexing. Most of these are statistically based models with computer learning.

As articles such as http://www.sciencedaily.com/releases/2008/09/080904151624.htm reveal digital photography is very effective in allowing records to be used in different locations at a fraction of the cost of photocopying.

Conclusion:

Some large companies are moving rapidly to a digital document collection. Government is moving more collections to digital format. ERIC, Education Resources Information Center, is moving its microfiche collection online as it obtains copyright permissions. The National Archives and Records Administration (NARA) is the Government agency that preserves and provides access to the U.S. Government's collection of documents recording the important events in American history. Their archival holdings number more than 10 billion pages of unique documents that are currently in the process of being digitized.Many banks have already been mostly digitized. Some hospitals have also made great progress. However, not much has been done in clinics and doctors' offices. And home collections have been relatively untouched so far. Our experience to date has lead us to believe that these smaller collections can be digitized effectively.