5th Annual Workshop on Technology for Family History and Genealogical Research
24 March 14, 2005

Presenter: Rick Laxman—Manger, Digital Imaging; Family & Church History Department, The Church of Jesus Christ of Latter-day Saints

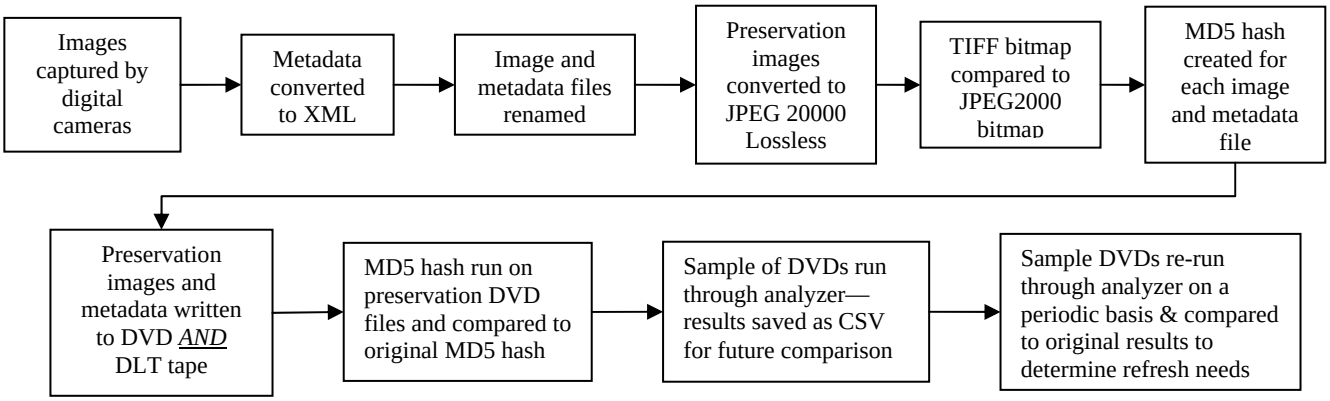
Title: Family and Church History Interim Digital Preservation Process

The challenge of long-term preservation of digital objects and metadata used for Family History research continues to be a significant problem in terms of methods, technology, and cost. Many academic, commercial, professional and private institutions are making efforts to solve this difficult problem.

The Family and Church History Department of the Church of Jesus Christ of Latter-day Saints is adopting an interim approach to preserving digital images and data. At present, we will be storing each image and metadata file on two sets of media: one on DVD and one on DLT tape.

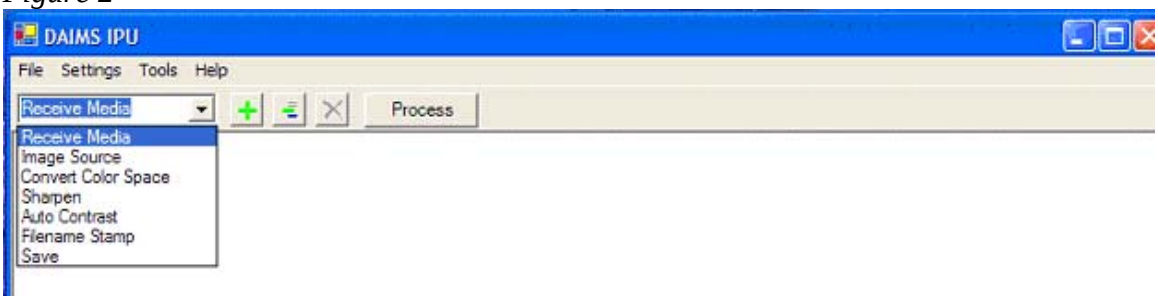
We designed a number of steps to prepare the files for storage and to make sure that the files are written without data loss. The diagram below illustrates those steps.

Figure 1



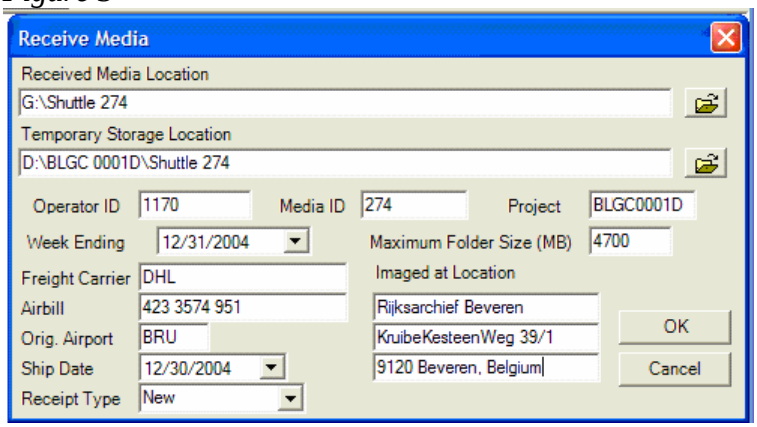
Folders or directories containing images and metadata are received from digital cameras located in archives throughout the world using the Image Processing Utility (IPU) (see Figure 2).

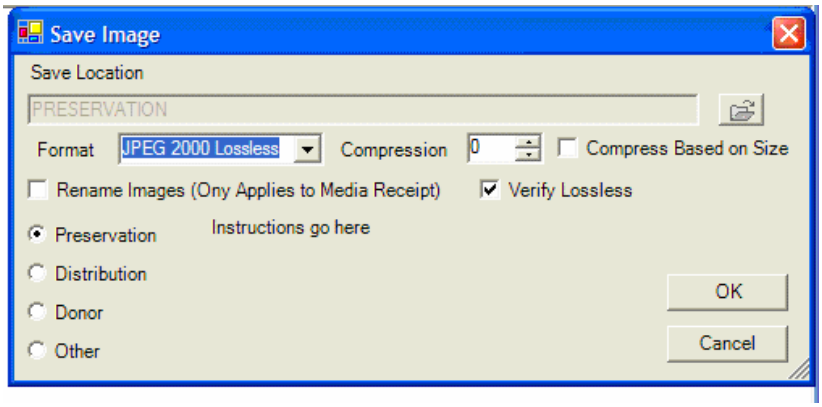
Figure 2



We developed the IPU to receive the folders, convert all metadata to an XML format, rename the images, and create an XML rename log as shown in figure 3.

Figure 3





We also create an MD5 hash value for each image and metadata file. An XML hash log is created to record the hash values. An example from the hash log is shown below.

```
- <Hash>
  <File>BLGC0002D_Code512-9999-000-00335-000_M_00001--.tif</File>
  <Hash>2ABB37D00C4D21FE891B8963146C3725</Hash>
```

When files are written to a DVD for preservation, hash values are created from the images on the DVD and compared to the values in the hash log to ensure that the files were accurately written to the DVD without data loss. Once the DVDs are written, a sample of the DVDs will be loaded on a DVD analyzer to measure the characteristics of the pits and lands that were burned into the DVD. The results of the analysis will be saved and used for comparison to other analyzer tests conducted at regular intervals in the future.

We considered both the MD5 algorithm and Secure Hash Algorithm (SHA-1) for creating the digital signature or fingerprint of the files. The MD5 algorithm creates a 128 bit (16 byte) message digest. The SHA-1 algorithm produces a 160-bit (20 byte) message digest. We felt the extra overhead in time to create the hash and additional storage space did not warrant the use of the SHA-1 algorithm at this time. The MD5 hash should provide the assurance that the images

were written correctly. In the future we plan on creating the MD5 hash at the time of image capture using the digital camera software. Table 1 shows the various hash standards/algorithms currently available.

Algorithm	Message Size (bits)	Block Size (bits)	Word Size (bits)	Message Digest Size (bits)	Security2 (bits)
MD5		512	32	128	60
SHA-1	< 264	512	32	160	80
SHA-256	< 264	512	32	256	128
SHA-384	< 2128	1024	64	384	192
SHA-512	< 2128	1024	64	512	256

Table 1: Secure Hash Algorithm Properties [1]
Note: MD5 information added by the author of this paper[2]

Another step in the receiving process is to convert the original TIFF, grayscale images to JPEG 2000 lossless for preservation. Our staff conducted image compression tests using J2K, PNG, GIF, WINZIP, and B-ZIP lossless formats to determine which format would provide the quickest method of compression, the smallest resulting file size, and the most stable format. Table 2 shows the test results.

Source	Byte Size	Time to Write in Minutes	Compressed % of Original	Reduction in File Size	Images Compressed Per Minute
J2K/ USING IPU – 200 DPI Microfilm Scanning	142,112,830	510	35%	65%	88
Digital Camera Capture	54,574,760	714	55%	45%	31
PNG/ USING IPU – 200 DPI Microfilm Scanning	201,785,603	510	49%	51%	176
Digital Camera Capture	72,469,653	346	74%	26%	64
GIF/ USING IPU – 200 DPI Microfilm Scanning	228,861,373	510	56%	44%	161
Digital Camera Capture	84,565,129	335	86%	14%	69
WINZIP/ USING IPU – 200 DPI Microfilm Scanning	200,134,094	510	49%	51%	207
Digital Camera Capture	73,467,078	247	75%	25%	191
B-ZIP/ USING IPU – 200 DPI Microfilm Scanning	65,595,354	510	39%	61%	16
Digital Camera Capture	59,956,671	1,030	61%	39%	23

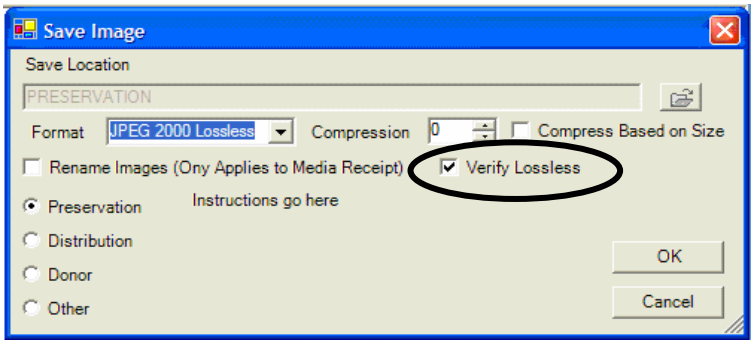
Table 2: **Image Compression Testing—Microfilm Scanning & Digital Camera Images[3]**

[1] National Institute of Standards and Technology. *Secure Hash Signature Standard (SHS) (FIPS PUB 180-2).* (page 3). Available online at http://csrc.nist.gov/publications/fips/fips180-2/fips180-2.pdf
[2] Touch, Joseph D. Performance Analysis of MD5. (Sigcomm '95, Boston MA). Available at http://www.isi.edu/touch/pubs/sigcomm95.html
[3] White, Herbert J. Compresion Testing, Final Report. (September 21,2004). Unpublished.

The test results indicated that JPEG 2000 lossless reduced images scanned from microfilm by approximately 65% and image from digital cameras by 45%. The difference in size can be attributed to the microfilm scanners cropping the images to the edge of the documents while digital camera images are cropped manually and leave a black border around the edge of the documents. We also found the JPEG 2000 format to be the most stable. Even though JPEG 2000 was not the fastest compression algorithm, the overall performance of the algorithm including the resulting file size, compression speed, and stability made it the obvious choice.

One final process step will be implemented to guarantee image integrity when creating a JPEG 2000 lossless image file from the original TIFF image captured with the digital camera. Since both formats are lossless, a bitmap to bitmap comparison can be made of an image in the two formats. We are testing a utility that compares the bitmap of the original TIFF image to the bitmap of the JPEG 2000 lossless version (see Figure 4). Using this approach we will be able to have confidence in the process of generating JPEG 2000 images for long-term preservation from the original TIFF images.

Figure 4



The department will be developing a process to monitor and recheck the preservation DVD and DLT tape media. We will test the media within one year of creation to determine if any problems exist with retrieving or reading the data or if there are problems with the media itself. Thereafter, we will complete testing on a regular basis. The testing will help us to know if we need to refresh the media by copying it to the same media type (e.g. DVD to DVD) or to migrate the files to a new technology when that technology is deemed viable.

As indicated previously the above process is temporary. The department believes that the future of preservation will be on servers and hard drives. We will be studying the possible methods for preserving images and data using server technology. The advantages of using server or spinning disk technology include

- Long-term cost will be lower
- Migration is achieved by adding new technology (servers, disks, etc.) to existing equipment when required to provide additional storage space, replacing failed equipment components, or as a schedule replacement as part of preventative maintenance
- Testing the viability of the files and media (disks and supporting components) is done in real time and automatically through built-in monitoring applications and algorithms
- The reloading of files in unnecessary to fill requests from Indexing, Description, etc. for images and metadata files or to recover, reprocess, and restore data to the hosting system