The FamilySearch Digital Process

Richard J. Laxman; FamilySearch International; Salt Lake City, Utah/USA

Abstract

FamilySearch International ("FamilySearch") is a non-profit organization that captures images and metadata from documents in archives worldwide and hosts them or facilitates archive hosting for individuals doing genealogical research. FamilySearch developed a "Digital Pipeline" to capture 40 million images and metadata per year from original records. The pipeline also includes scanning 65 million images in 2008 from the 2.4 million roll microfilm collection. These images and metadata are processed, cataloged, indexed, hosted and preserved. The organization has developed hardware and software such as a digital camera and microfilm scanning systems that are used in the Digital Pipeline. An Internet indexing application known as FamilySearch Indexing is available to volunteers to index images from their homes. FamilySearch develops and implements internal standards for images and metadata. These standards are presented to national and international standards organizations for consideration as national and international standards.

This paper will present the various processes within the Digital Pipeline and the tools and standards which were developed to facilitate those processes. It will explain the cooperative, ongoing efforts of FamilySearch with archives and record repositories worldwide. Details on how FamilySearch works with third-party affiliates to capture, index and host images for archives will also be outlined.

FamilySearch

FamilySearch, formerly the Genealogical Society of Utah, is a non-profit organization that gathers records for the purpose of genealogical research. Today this is done using a combination of digital camera systems and microfilm camera systems developed by FamilySearch. This paper will discuss the digital process used by FamilySearch to capture, scan from microfilm, process, prepare, deliver and preserve digital images and associated metadata.

These records are found in various states of organization and deterioration and in various formats such as loose papers and bound material. FamilySearch has a program to digitize and preserve these records.

Digital Pipeline

FamilySearch developed a Digital Pipeline to create digital images from microfilm scanning and from original documents captured with digital cameras. Included in this pipeline is the receipt and processing of images, waypointing and indexing of images, hosting and distribution of images and metadata and preservation of images and metadata. Quality and standards play a prominent role throughout the Digital Pipe line. Figure 1 shows the Digital Pipeline developed by FamilySearch.

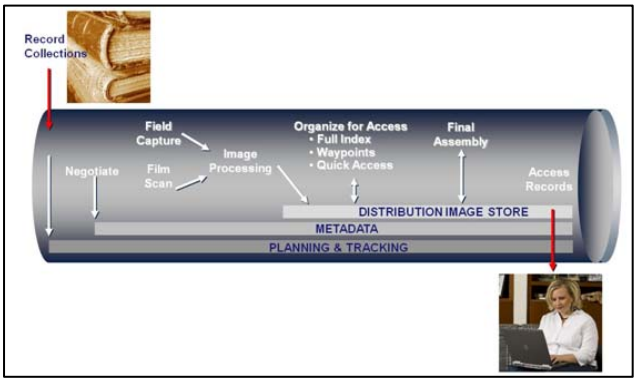

Figure 1. *FamilySearch Digital Pipeline*

Negotiations

FamilySearch negotiates with archivists and records repositories worldwide for the rights to image records of genealogical research value. These records are processed and posted on the FamilySearch Web site for members of FamilySearch and others to search for their ancestors. The archives and record custodians receive copies of the images in return for allowing FamilySearch to digitize their records. FamilySearch may also work with other parties in joint ventures to digitize the records of an archive. FamilySearch also negotiates with archives to scan records that FamilySearch has previously microfilmed. The purpose of these negotiations is to obtain permission to scan those microfilm records for hosting on the FamilySearch or the archive's Web site.

Microfilm Scanning

FamilySearch has developed a new microfilm scanning system. The organization developed software and modified hardware to be able to scan a single roll of microfilm to one large image file known as a ribbon. Using specialized software the individual frames are selected from the ribbon and outlined for cropping. A quality evaluator reviews cropped frames for accuracy and makes adjustments as necessary. The evaluator also reviews the images for contrast quality and adjusts the contrast to improve readability of the image as necessary. See Figure 2.

Figure 2. *Microfilm scanning quality assurance*

Images & metadata from microfilm scanning are forwarded to the next process steps: Indexing and Waypointing. The original microfilm is considered the preservation copy and is preserved according to established international standards in an environmentally controlled, secured facility. See Figure 3. It is less expensive and more efficient to rescan and link the index and waypoint data to the images than to preserve the digital images.

Figure 3. *Underground storage facility for microfilm and digital media*

Digitization of original documents

FamilySearch developed a digital camera system for digitizing original documents in bound and loose formats. These cameras are located in the archives and libraries throughout the world. FamilySearch engineers developed the camera stand and software used for the digital capture system. Industrial grade digital cameras are integrated to capture documents of various sizes and formats. Currently the camera systems use 11 and 16 megapixel cameras (CCD array) in a mounted overhead system for digitizing documents. See Figure 4. FamilySearch is now developing a 50 megapixel camera system and a color system. A CCD array camera is used for the rapid capture of documents.

Figure 4. *FamilySearch digital capture system*

Due to a lack of portable camera systems that would provide for the capture of document metadata, FamilySearch decided to develop its own digital camera system. This effort first began in 1998. The first digital camera project was carried out as a joint project with the National Archives of Scotland. The preliminary setup of the project began in 2000 and the digitization of the documents began in 2001. FamilySearch now has over 160 digital camera systems operating in archives worldwide.

Features of the current digital camera system include the automatic importing of the listing information prepared by the organization's collection specialists and by field representatives. The information is stored in an XML format. See Figure 5.

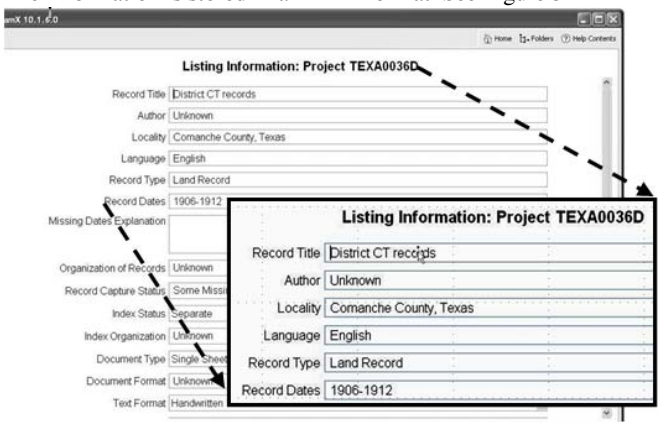

Figure 5. *Listing information imported into digital camera metadata store*

Other features include a semi-automated method for manually focusing the digital camera. This method uses a target and a focusing algorithm to help the camera operator accurately focus the camera. As the camera is brought into focus two lines on the chart line up and a checkmark appears when the camera is accurately focused. See Figure 6.

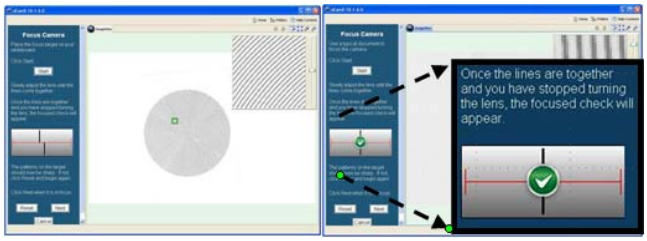

Figure 6. *Camera focusing algorithm*

The camera system also includes an automated process for determining resolution. FamilySearch uses a resolution measurement known as pixels-per-line-segment (PPLS). The required resolution for a given document is determined by imaging a sample of a volume or logical grouping of loose records which contains an example of the thinnest line segment of a character within the volume or documents. The camera system measures the number of fully and partially blocked pixels across that line segment. The required minimum number of a combination of partially and fully blocked pixels is three pixels across the line segment. If a line segment meets the standard, the segment turns green otherwise it turns red. See Figure 7.

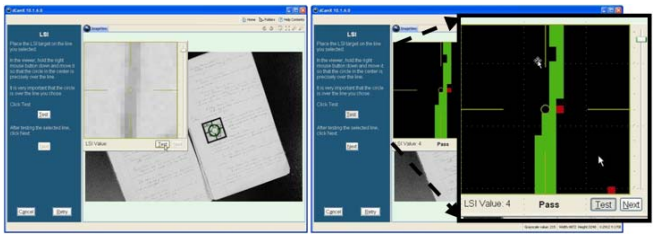

Figure 7. *Pixels-per-line-segment measurement*

The FamilySearch original image format is TIF, grayscale uncompressed. Images can be captured one or two pages at a time. The software allows the odd numbered pages to be captured first and subsequently the even numbered pages. The application will then place the pages in the correct contextual order. Within the digital camera software is a quality evaluation module that requires the digital camera operator to review the quality of each image after capturing a group or volume of documents. An MD5 hash log is created for all of the images captured. The hash is used for image verification throughout the pipeline. The image capture screen is shown in Figure 8.

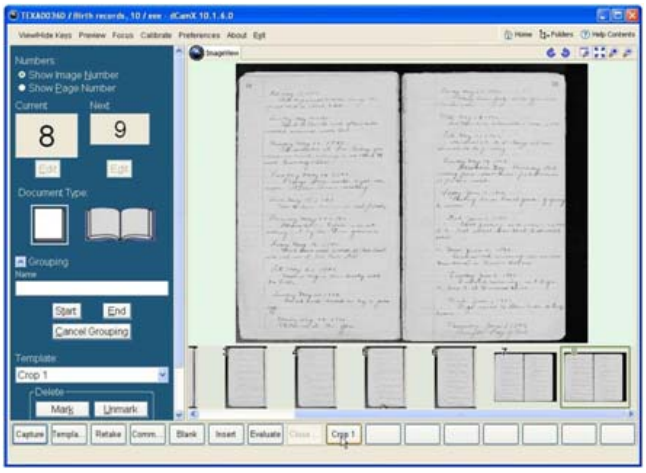

Figure 8. *The capture and evaluation of images*

Image Processing

Once the images are captured they are transferred to an external eSATA hard drive. Every time images are transferred, the systems check the MD5 hash for each image and metadata file. The hard drive is shipped to FamilySearch in Salt Lake City, Utah where the images are processed. Derivative images are created from the original TIF images for the next steps of the process:

- Lossless JPEG 2000 images are created and written to preservation media.
- Generally the original grayscale TIFF images are provided to the record custodians as the donor copy. The metadata captured when the images were digitized and subsequently added is also included with the donor copy. The donor may also receive a copy of the images at the FamilySearch cost. Donor images and metadata are sent to the donor on an external hard drive or DVDs; external hard drive is the preferred method.
- Compressed JPEG images are created for the purposes of waypointing, indexing and hosting on the Internet.

Preservation

The current interim preservation method is to write images to a SATA hard drive and two LTO2 magnetic tapes. The files on the hard drive are locked on the drive by changing file attributes. The hard drive is stored in a static blocking bag and placed in a graphite coated box. The graphite coating stops EMF fields from reaching the hard drive and erasing or corrupting its data. Cyclical Redundancy Checking or CRC logs are written and verified for every 2 K bytes of data on the LTO2 tape. The hard drives and tapes are sent to FamilySearch's underground storage facility for preservation shown in Figure 9. The drives and tapes are monitored for degradation of media or the bits.

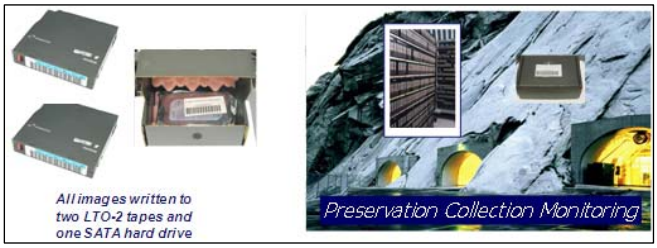

Figure 9*. Preservation of LTO2 tapes and hard drives*

Authorities

FamilySearch authorities consist of expert information that is integrated into a database in order to help non-experts search for names, places, or dates. For example, a name authority for "Margaret" gives spelling variations, abbreviations, and nicknames for that name. A place authority for Brasov, Romania, lists variations of the city's name, its political and ecclesiastical jurisdictions, and a history of boundary changes. Such information enables anyone to easily access expert information without having to be experts themselves. Examples of authorities information showing boundary and names changes is shown in Figure 10.

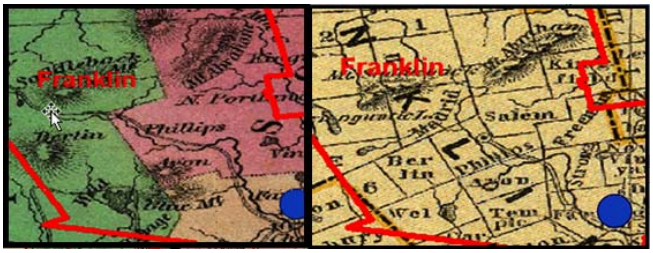

Figure 10*. Authorities information – boundary changes*

Waypointing

Waypointing is a method for providing access to collections of digital images without providing a complete index to each record in the collection. Using the waypointing application shown in Figure 11, FamilySearch associates place specific markers at points in the collection of images that allows users to more easily browse that collection. For example, a collection of birth records for the state of Utah could contain markers to allow users to select the specific county and year to search. Users would then browse through the images for only that county and year. Waypointing improves the efficiency of browsing records, making them accessible without having to index every record in the collection.

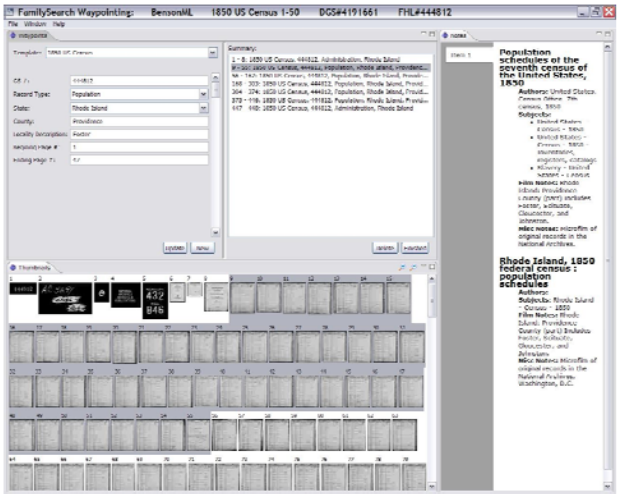

***Figure 11**. Waypointing application*

FamilySearch Indexing

FamilySearch gathers genealogical and historical records from around the world and converts them into digital images which are stored in its online system. The key life events of billions of people are available on these records and are preserved and shared through the efforts of volunteer indexers. Using the FamilySearch online indexing system, volunteers worldwide from FamilySearch and from other organizations, usually having interest in a specific collection of records, are able to quickly and easily transcribe the records—all from the convenience of their homes. The indexes and associated images are then posted for free at familysearch.org.

Volunteers download a batch of images to their computer and transcribe the highlighted information in about 30 minutes per batch. As a volunteer moves from field to field in the indexing software, the area on the image to be entered is highlighted. After the volunteer finishes indexing a batch of images and reconnects to the FamilySearch site the indexes are uploaded and the images are deleted from the volunteer's computer. Details can be found at http://www.familysearchindexing.org/home.jsf. Figure 12 shows a screenshot from the FamilySearch Indexing application.

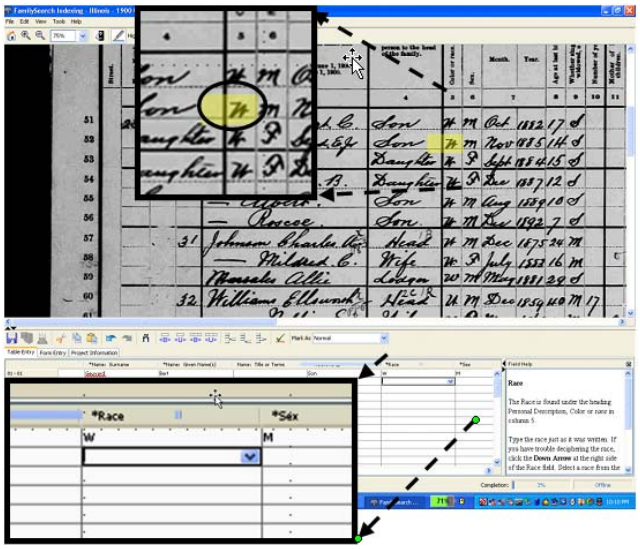

***Figure 12**. FamilySearch Indexing application*

Research Guidance

Research help is provided for collections that are posted on the FamilySearch Web site. The site provides information about the collection, how to use it, contextual information and the history of the collection. Figure 13 is an example of the research help provided at the FamilySearch Web site.

***Figure 13**. Research Guidance example*

Hosting and Record Access

JPEG derivatives of the images captured from original documents or scanned from microfilm are combined with authorities information and waypointing data or indices in preparation for hosting images on the FamilySearch Web site. Record Search is the online product through which the public can search and view FamilySearch's digitized genealogical records. Many of these records have been indexed by FamilySearch Indexing volunteers. Record Search also links to images and indexes from third party affiliates. Record Search is currently available as a pilot from FamilySearch.org or directly at Labs.familysearch.org. Figure 14 displays the portal to reach Record Search.

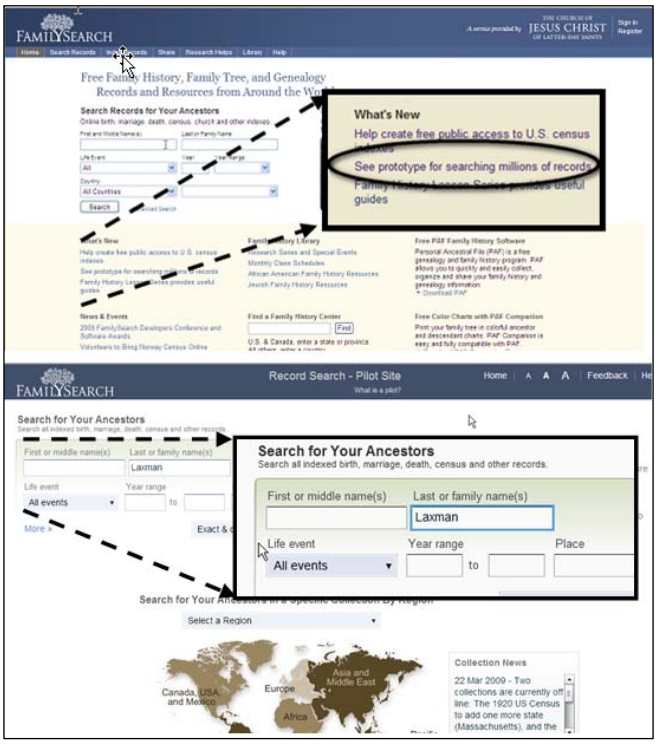

***Figure 14**. Record Search portal*

Figures 15, 16 & 17 exhibit the results of a name search conducted at the Records Search site.

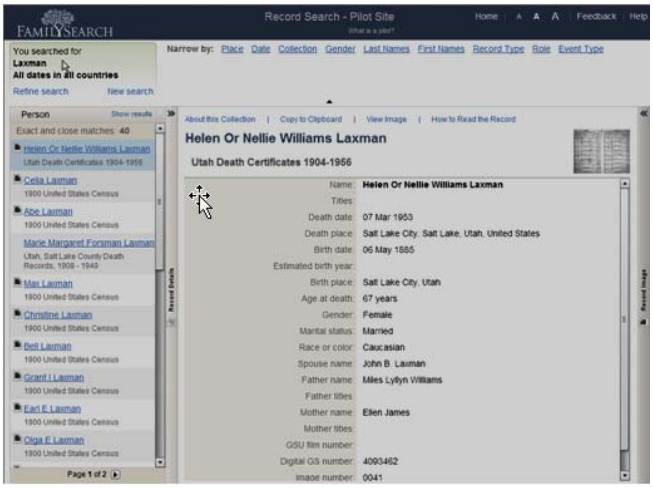

***Figure 15**. Name search results and metadata*

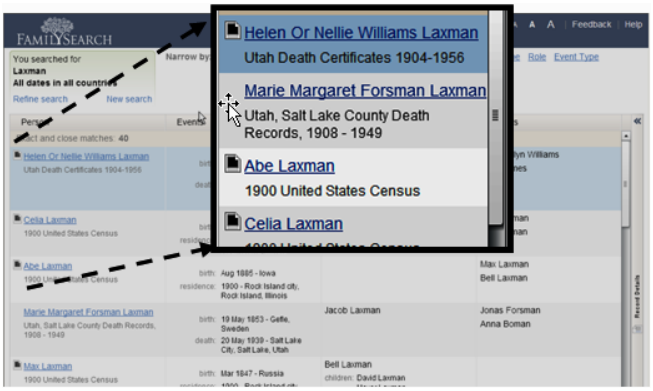

***Figure 16**. Names in search results*

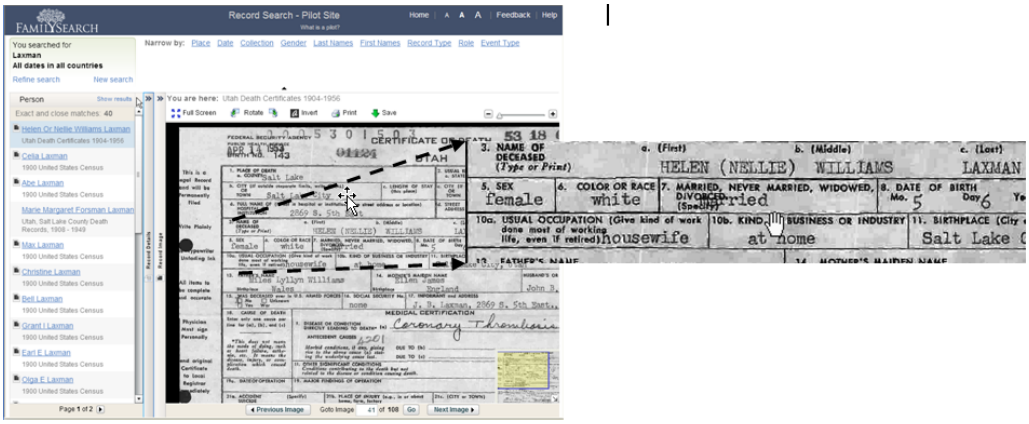

***Figure 17**. Image associated with name search result*

FamilySearch Standards Involvement

FamilySearch is a pioneer in national and international standards development for both microfilm and digital imaging technology. FamilySearch representatives served on and chaired the standards development committees for the Association for Information and Image Management/American National Standards Institute (AIIM/ANSI) and the International Organization for Standards (ISO). Today FamilySearch representatives are involved in the following standards committees:

- AIIM/ANSI standards program
  - Chair, C24 Electronic imaging
  - Member of other committees
  - Member of Standard's Board
- ISO standards program
  - Member of TC 171, Document Management Applications

FamilySearch also develops internal standards such as the *FamilySearch Digital Imaging Specification*. This standard is used to establish and enforce specifications for image format, quality and delivery. Internal standards and specifications developed by FamilySearch have been and will continue to be proposed for consideration and adoption by national and international standards organizations for use as industry standards.

About FamilySearch International

FamilySearch International is the largest genealogy organization in the world. Millions of people use FamilySearch records, resources, and services to learn more about their family history. To help in this great pursuit, FamilySearch has been actively gathering, preserving, and sharing genealogical records worldwide for over 100 years. FamilySearch is a nonprofit organization sponsored by The Church of Jesus Christ of Latter-day Saints. Patrons may access FamilySearch services and resources free online at familysearch.org or through over 4,500 family history centers in 70 countries, including the main Family History Library in Salt Lake City, Utah.

Author Biography

Richard Laxman championed digital technology and imaging processes development for FamilySearch. He developed the first digital camera system for the organization and today manages its digital data processing center. Mr. Laxman received a master's degree in Business Information Systems (2005) from Utah State University. He serves as committee chair of C24, Electronic Imaging in the AIIM standards program, project editor for an ISO metadata project and serves on the AIIM Standards Board.