Metadata: Key to High-Volume Access to Records

Emily S. Schultz and Paul D. Abbott; FamilySearch; Salt Lake City, Utah/ USA

Abstract

FamilySearch has the goal to preserve and publish the worlds vital records. In a presentation last year, Archives members were shown how FamilySearch's dCam-X tool can speed up and improve the process of capturing digital images. The next step is to publish the records online with useful metadata, making them accessible to the public as quickly as possible.

FamilySearch is committed to digitally publishing the records it captures. The key to doing this on a large scale is identifying, gathering and processing metatdata about the content of the records being captured.

Introduction

FamilySearch is all about preserving and publishing the world's vital records. These collections are gathered in from the great archives of the world along with being gathered from remote villages on the plains in Africa, from the mountains of Latin America and from the islands of the South Pacific. FamilySearch works with records that are as diverse as the cultures, locations and people of this planet. And yet, as diverse as people and cultures are, there are common events and relationships that define the human experience. There are also innate desires to preserve the history of these common events. The collections of these historical events are the vital records FamilySearch gathering and publishing.

The purpose of this paper is to provide an overview of the processes, tools and standards used by FamilySearch to use metadata gathered to publish to the web hundreds of millions of images. The reader will see that the key to managing high volumes of data is having the right metadata in place from the beginning.

Traditional Cataloging vs. Digital Indexing

The needs of patrons have rapidly changed over the past decade. Once they were content to go to a library, search through the catalog and then browse through the library to view the physical items they were interested in. Now, because of the availability of easy information on the internet, they want instant access to archival records. Patrons today want to be able to click on a record, and instantly see the image associated with that record.

FamilySearch has excelled at capturing the world's records on microfilm, and making them available around the world. FamilySearch has captured 2.4 million rolls of microfilm from nearly 200 countries. There are over 4,500 Family History Centers around the world that allow access to these records. Maintinaing the catalog for these records has been critical for allowing partons to find the information they are looking for. However, as FamilySearch moves to digitally capturing hundreds of millions of new images and converts the over two Billion images in our microfilm collection, maintaining a manual catalog is no longer feasible.

A Patrons experience with accessing records on the web will vary depeding on the records they are looking for and how much work has been put into making the records digitally searchable. For printed books, every word can be scanned, read and indexed by computer systems. However, most of the vital records being captured by FamilySearch are hand written and difficult for computer systems to read. Therefore in order to make records easily available, processing the information on the record is required, this usually means armies of people maticulously typing in key-word informaiton into a database. As shown in Figure 1, FamilySearch provides 4 different levels of processing records: Linked Family Data, Full Name Indexes, Light Indexing and Field Express. The pyramid indicates the relative number of records being published by FamilySearch and the level of manual processing done. The amount of processing performed will vary depeding on the demand for the records and the richness of information on the records.

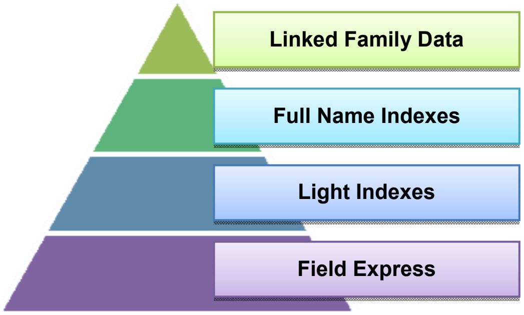

Figure 1: Processing Pyramid

Linked Family Data

The ideal for FamilySearch, and for its users, is to show the worlds' records and images in the context of families. Family names would be transcribed from the original records, and then linked into family groups. A patron searching a name would find a link to take him to not only that person, but their parents, siblings, and children. Each person would have several corroborating original documents, in digital format, that were available as sources for verifiying the information.

FamilySearch has done some work with Family Linking. However, it is very time consuming and requires both sophisticated algorithms as well as very skilled human analysis and auditing.

Full Name Indexes

The next best type of access is a full-name index to the records. This is something that FamilySearch has also pioneered, and excelled at. The FamilySearch Indexing software is an amazing tool to allow volunteers to capture names and other important data points from digital images of original records. Currently, FamilySearch has over 100,000 volunteers from all over the world helping out with indexing. Current indexing projects can be found at:

http://indexing.familysearch.org

The result of this indexing allows patrons to freely search a name index, and then go to the image of the original record. He or she can then extract other information from the record, especially those details that will allow linking into families.

Light Index or Browse

Even though FamilySearch is indexing at a very impressive rate, and have plans to grow that rate even further, it still cannot index fast enough to capture all the records that it captures and preserves.

So, another form of indexing, a light index, will allow FamilySearch to index only the place name, event type, and date. This does require more effort and research on the patron's part, but will result in a similar experience to the former library catalog – a set of records that a patron can search. Instead of going to the library to look at microfilm, he can look at the records online.

But again, this light indexing does require manpower, and FamilySearch does not have the required number of skilled people to keep up with the images captured.

Field Express

Because it is impractical to manually review and catalog every set of images, FamilySearch has developed a new program called Field Express. This level of access allows for browesable images based upon "information tags" captured at the time of filming. With Field Express, the metadata is attached to groups of images as they are captured, instead of by catalogers after capture. This successful program allows patrons to browse the records right after they are captured digitally.

FamilySearch will allow immediate access to records with Field Express. As resources allow, FamilySearch will hope to move up to different levels of access with each record set, and allow the users to link them into families.

Field Express clearly offers the best path for the majority of images. The images can then be made available online for viewing nearly as quickly as they are captured

Field Express

Field Express Publishing Process

The first step to attach the metadata to the images is to create a FamilySearch project, a contractually afreed upon list, or listing, of the collections to be captured. A tool is used to input the inventory or catalog information from an archive, showing the content to be captured.

The key pieces of metadata to be captured include:

1. Locality to the lowest jurisdiction in the record set (Examples: Panama City, Washington County, Saint Paul Diocese)
2. Record Type (Examples: Marriages, Deaths, Probate Records, Military Records)
3. Dates (Examples: 1850, 1890-1899)
4. Volume Numbers (Examples: Packet 0923, Volume 26, Container 35)

Organization of Records. Clearly it is very important to be aware in advance of the organization of records in the archive. Many records are organized by date, by city, by packet number, etc. Knowing the hierarchy of the metadata is essential.

Currently, the dCamX software does not allow ingesting of existing registers or inventories. In the near future, this feature will streamline this step even further. By ingesting an existing inventory, not only is the archive's own metadata being leveraged for use, but it is an important quality step that will keep us from missing records or from mis-attributing records.

Camera Capture using dCamX

Once the software has the basic metadata structure entered, the information is sent on a shuttle to the camera operator. The camera operator then captures the images, and assigns the metadata as appropriate to each folder. Camera operators can make corrections to the metadata as he goes along, but too much of this will slow down production

Processing

After the metadata is captured with the images, the files are sent on a shuttle to FamilySearch. The images are sent through a quality audit step, and the data follows a similar quality audit step, to ensure it is publishable. The images are sent to a central image distribution store. The metadata is sent to a data store, where it can be accessed by the publication software. The metadata contains links to each image.

Publication

Currently, FamilySearch publishes new content weekly. The site contains records for 67 countries, and over 550 collections. Each week FamilySearch publishes both new collections, and updates to existing collections. The Field Express images are posted as browseable collections, using the very metadata that was input at the beginning of the process.

Current Status

FamilySearch is just a few months into the Field Express pilot. Over two million images have been published, from several dozen collections.

Some of the learnings already for the Field Express pilot include:

1. The importance of standards. Without standard metadata, the browse will be cluttered, and hard to navigate
2. Feedback loop. One of the best ways to give feedback to a camera operator about the metadata that they are capturing, is to show them how the metadata looks when it is published as-is online.
3. The importance of having a good listing to start with. Basing the image capture on good existing metadata results in cleaner, more accurate access to records

Future Plans

A future state would allow us to ingest existing indexes and match them to records as they are filmed.

Conclusion

FamilySearch is capturing tens of millions of images per year of the world's vital records. As the rate of capture and publication increase, with a goal of a million images per week, the importance of metadata is key to providing good access. The rate of publishing a million images per week is but a starting point. By expanding tools and improving process to meet the global needs of archives and record donors, FamilySearch hopes to eventually be able to grow and capture vital records at the rate they are actually being created.

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 Boards
- 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

FamilySearch 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 or 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

Emily S. Schultz is currently a project manager at FamilySearch, where she has worked for sixteen years. She is the former manager of the Family History Library Cataloging Department, and has over 20 years of library experience. She received her MBA from Weber State University and her BA from Brigham Young University.

Paul D. Abbott received his BS is Mechanical Engineering from Brigham Young University and his Masters in Engineering Management (MEM) from Brigham Young University. He is currently responsible for defining and developing the next generation of digital tools and processes needed to rapidly gather and publish collections at FamilySearch.