Images Online at Ancestry.com: Technology and Workflow Overview

Extended Abstract, JGG, 27 February 2001

When the current census phase is completed, Ancestry.com's Images Online project will have posted over 12 million images on the World Wide Web in searchable and/or browseable form. This paper will describe the technical, managerial, and genealogical challenges involved with the project.

Background

A great portion of the data used in genealogical research originates as the result of the activities of governments or other large organizations. Until recently, therefore, most actual genealogy on the Internet has been based on the use of secondary sources compiled from or based on primary governmental or other organizational records.

Correspondingly, most of the names in Ancestry.com’s online databases have come from secondary sources—indexes, extracts, etc. Researchers are encouraged to seek out documents and records that are most closely associated in time and place with the events they record. For this reason, many genealogical and historical researchers seek to go beyond the limited information currently available in an electronic index and view or obtain copies of the original records from which these indexes were created.

The Images Online™ project is based on the ability to deliver online a digitally scanned copy of an original document to a customer. The project allows the genealogist to get a digital copy of an image much more quickly and cheaply than through conventional means. Records custodians and archivists, such as the National Archives and Records Administration, generally provide limited service with extended waiting times. In many cases, these large organizations have no charter or organization to fulfill the needs of a growing genealogical customer base.

The U.S. Federal Census

The U.S. Federal Census was selected as the record group for this Images Online project for a number of reasons: a) it is generally acknowledged that federal census records are the most valuable set of genealogical records for U.S. research; b) census records cover a large time span from 1790 to 1920[1] and enumerate the entire U.S. population at the time they were taken; c) much of the census records, especially 1850 and later, contain great detail about families—much more than has been generally available from census indexes alone; and d) census records created by the federal government are in the public domain.

Finally, as is the case with most large collections of genealogical records, most researchers do not have convenient access to a nationwide collection of census records. Accessing such a collection might require a trip to a university or research library, a regional branch of the National Archives and Records Administration (NARA),

[1] The U.S. Federal Decennial Census has been taken every ten years from 1790 to 2000, though only those censuses at least 72 years old have been made available to the public.

or an LDS Family History Center, where individual rolls of federal census microfilm must be ordered from the main library in Salt Lake City.

Image Acquisition and Enhancement

Any ambitious project must begin with good source information. The images in Ancestry.com's Images Online census project are scanned from second-generation microfilm acquired from NARA. The film is cleaned and loaded into a high-resolution, high-speed microfilm scanner. The film's quality and photographic exposure determine the initial digitization settings. This scanner and the scanner operator then scan each frame of the microfilm as a separate image in 256 levels of gray. Each frame creates a 5-8MB TIFF file.

Obviously, to the genealogist, the quality of the scanned images is of major importance. Census images available from the Internet are not valuable if they are not scanned well enough to allow them to be read. In addition to quality control at the scanning station itself, each census image undergoes a detailed clean-up process. Once again a human operator views a scanned image, which has been automatically sharpened, de-skewed, adjusted for contrast, etc.

If the resulting image is not an improvement, or if the image is still not high enough quality to use in genealogical research, further cleanup processes are applied to the image. In worst cases, the image is rescanned from microfilm.

Overall, the most challenging requirement of the project is production capacity. In order to satisfy market demand for these valuable census images, the images must be scanned and processed at high speed. At maximum production the process generates up to 100,000 images a day.

Image Compression and Delivery

Not including the 1930 U.S. Federal Census, which will become available in April 2002, the entire census collection numbers over 10 million images and contains information on 450 million individuals. Though these large original files have been preserved for possible future use, the images available at the Ancestry.com Web site have been compressed both to reduce storage requirements at Ancestry's hosting facility and to ease the bandwidth burden on the genealogical customer, many of whom still use 28.8Kbps modems for Internet access.

Ancestry selected the *Multi-resolution Seamless Image Database (MrSID)* format to display its census images over the World Wide Web. This choice was motivated by the following factors:

- *Extreme Compression with Minimal Visual Data Loss*. Using wavlet-based compression algorithms, the MrSID data format achieves extremely impressive compression ratios (near 25:1 on most census images) without an excessive loss in image quality. In fact, Ancestry's 8-bit grayscale images are roughly the same size as black and white census images available from competitors.

- *Sub-Region Extraction*. The MrSID technology optimizes bandwidth by only delivering portions of the image to the end user (rarely will an individual look at an entire document at its maximum resolution). MrSID only delivers those pixels to the end user which are needed to display whatever section is currently requested.

- *Single-Source Imaging*. MrSID images can be retrieved by an Internet browser in a variety of sizes/resolutions—the same images can be used to show browseable thumbnails as well as the detailed scans examined at high zoom for genealogical purposes.

Once scanned, cleaned, and compressed, the census images are made available to customers through their Internet browser at the Ancestry.com Web site. Many images will be linked to a pre-existing census index at launch, allowing users to search for individuals by name and filter their searches by locality and census year. For those censuses which do not yet have a corresponding digital name index to their contents, Ancestry.com has provided a browse structure wherein users can select census images by census year, U.S. state, county, and township and/or enumeration district.

Conclusion

Ancestry.com's Images Online census project has been extremely successful. Most of the 1920 and 1900 census have now been posted, and new images are coming online at the rate of several hundred thousand images per week. The collection is available to customers by separate subscription, which has been extremely popular.