

Management and Distribution of Historical Document Images

Rayman D. Meservy, Stephen W. Liddle, and John W. Welch
Brigham Young University
Provo, Utah, 84602

The Church of Jesus Christ of Latter-day Saints has some 50,000 CD's containing images of handwritten journals, letters, financial notes and many other interesting documents relating to the early Utah pioneers. Each image is currently in JPEG format, most with a larger image and also a medium-sized image. BYU Studies is working with the Church to determine the best methods of distributing these images to libraries and interested researchers in an efficient and economical manner.

The sheer number of document images and the related CD's poses a significant problem. The raw capacity of 50,000 CD's at 680 megabytes per CD is 34 terabytes (TB). Even assuming a low utilization rate of 60% on average, this still amounts to at least 20 TB of raw data.

BYU Studies is currently working on issuing sub-collections of these historical documents. For example, the Charles Rich Diary collection currently resides on six partially filled CD's. We can publish this particular collection by using a lower JPEG quality level (we can compress images by an additional factor of 4:1 or 5:1 without reducing the legibility of documents). Thus, one possible publication model is to distribute smaller collections in lower resolution on CD, and provide on-line access to higher-resolution images only for those who really need the very best resolution possible (and can afford the required network bandwidth).

There are several alternatives we should consider. First, high-quality document images could be maintained on one or more servers with thumbnails images and indexes widely distributed to the public. A server-based approach raises questions of how the database should be distributed and replicated. Also, what kind of compression should be used? Would multi-resolution browsing be appropriate for these kinds of historical documents? Multi-resolution browsing is similar to the idea of an interlaced GIF image, where incremental data transfer provides progressive levels of detail. For example, the Alexandria Digital Library Project uses wavelet decomposition to segment an image into parts that can be easily browsed at various levels of resolution [1]. Just-In-Time-Browsing is another project that provides a multi-resolution approach to image browsing [2]. These and other techniques should be considered for any server-based proposal.

Second, perhaps JPEG is not the best image compression technique for these historical documents. More highly compressed imaging methods would make distribution more feasible. Another approach to increase the effective compression ratio is to use a higher-density storage medium. For example, there are several consumer-level DVD recording formats that are now readily accessible (DVD-R, DVD-RW, DVD-RAM, and DVD+RW). These or other similar high-volume media would aid distribution.

Third, transmission of many images could be avoided if the documents were available in textual format. Common methods for specifying transcripts of hand-written manuscripts are available. If we had such transcripts for all the images, not only would we save transmission bandwidth, but also the database would be much easier for researchers to use and search. So what is the best way to construct such transcripts (or other indexes if transcripts are not used)? Is outside volunteer help available?

In our presentation we will describe the current status of this project, discuss design alternatives, and solicit discussion and suggestions regarding future activities.

References

- [1] B.S. Manjunath, "Image Browsing in the Alexandria Digital Library (ADL) Project," *D-Lib Magazine*, August 1995.
<http://www.dlib.org/dlib/august95/alexandria/08manjunath.html>
- [2] D.J. Kennard and W.A. Barrett, "Just-In-Time Browsing for Digital Microfilm," *Workshop on Technology for Family History and Genealogical Research*, March 29, 2001. <http://www.fht.byu.edu/workshop01/fht2001prog.php>