The Mormon Diaries Project

Frederick Zarndt
CTO
iArchives
frederick.zarndt@iarchives.com

Russell Black
Senior Software Engineer
iArchives
rb@iarchives.com

Scott Eldredge
Digital Initiatives Program Manager
Harold B. Lee Library
scott_eldredge@byu.edu

Jeri Jump
Project Manager, Scholarly Publications
Harold B. Lee Library
jeri_jump@byu.edu

Abstract

This paper describes the motivation, design, and use of a workflow management system and application software to transcribe diaries of Mormon missionaries written in the 20th and late 19th centuries. The transcription system is a joint, on-going effort of Brigham Young University's Harold B. Lee Library and iArchives.

Motivation

Transcription of manuscript materials has long been a staple process for library special collections. Recent technological advances have opened the door for new approaches to transcription, moving it past typewriters and word processing, and provided convenient avenues for access to transcribed materials. Participation in recent projects has allowed the staff at Brigham Young University’s Harold B. Lee library to experiment with several new approaches to manuscript transcription and has opened the door to a variety of possibilities for the future.

In 2002, with support from a Library of Congress grant, and in partnership with the University of Utah, Utah State University, and the Utah State Historical Society, the Lee Library made available a collection of 49 pioneer diaries and associated materials (http://overlandtrails.lib.byu.edu/). The diaries were transcribed by library staff using, for the first time, digital images of the diary pages as working copies. Also for the first time, the transcriptions were tagged, using XMetal software, in Extensible Markup Language following the document type definition for TEI Lite. Once the markup was complete, Library staff were able to associate the marked up text, including tagged fields with digital images of the pages, and import the data into CONTENTdm, a digital library collection management software package, for searching and display. The marked up text was also exported to easy-to-read PDF files that were likewise loaded into CONTENTdm. The result was a highly searchable and easily navigable collection of handwritten diaries.

An evaluation of the processes involved in creation of the Overland Trails Collection led to a search for more efficient ways to transcribe diaries and other manuscript materials. The volume of valuable manuscript material located in the Lee Library's collection and others around the world is enormous, but experience has demonstrated that current models could be never be scaled to successfully complete large-scale transcription projects.

The Mormon missionary diary collection is almost 10 times larger than the Overland Trails collection, consisting of approximately 70,000 pages. Funding would not cover the difficult and time-consuming process of using XMetal to transcribe, encode, review, and provide name normalization and tagging. A transcription and encoding software application and workflow management system appeared to be what we needed for this and similar future projects. Fortunately, iArchives had recently completed a prototype of a similar system for the Church of Jesus Christ of Latter Day Saints, the Internet Indexing System, and was willing to adapt it for transcription projects.

Initially students will transcribe the diaries, but we expect to expand the scope of the project so that community volunteers can also participate. In order to efficiently facilitate community participation, the data flow must be centrally managed using web-based services and protocols such as those provided by iArchives for the Internet Indexing project.

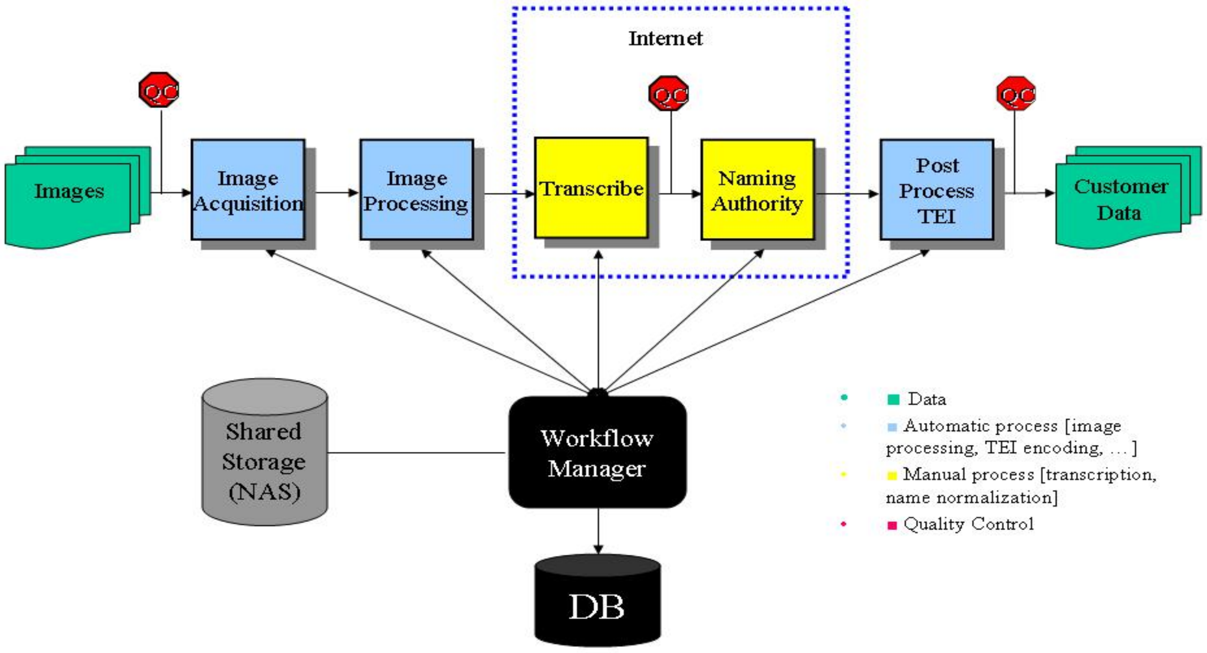
Technology

Performance improvements in Optical Character Recognition (OCR) technology, the falling cost of storage, distributed processing, and increased network speeds have made large digitization projects of machine print practical. However, accurate algorithmic recognition of unconstrained hand-written documents remains problematic, if not impossible (cf. On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey, Rejean Plamondon and Sargur N. Srihari in IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 22, No 1, Jan 2000).. Manual transcription, the alternative to algorithmic recognition of hand-written documents, can be error prone and expensive, but using a well-designed workflow management process and quality controls, both error rate and expense can be reduced.

Since summer 2000, iArchives has developed a configurable state-driven workflow management system (WFM), used by iArchives and several foreign partners to digitize text images for customers ranging from small community libraries to large university research libraries to corporate archives. The WFM uses a combination of commercial, open source, and other readily available, inexpensive (free) software tools such as Linux, Postgresql, Java, Xerces, LDAP, etc. The WFM manages and distributes work units -- a group of images -- to a series of project specific applications, each of which is associated with one or more states in the digitization process. The applications, which may be automatic (no human intervention necessary) or manual (user driven), request work units of a type suitable for the application from the WFM through an application framework. When the application completes a work unit, data derived from it is put in the location that was specified by the WFM, the WFM is notified, and another work unit is requested. Because the applications may run on computers either inside or outside the firewall protecting the WFM, the system has been designed so that it is simple to add additional processing nodes.

Each processing node registers its capabilities with the WFM: An automatic processing node registers through a Windows service or Linux daemon and tells the WFM via an HTTP-based protocol the process types – image processing, post processing, etc. -- supported by the node. For the Mormon diaries workflow, automatic processes include image processing and post process.

A manual processing node registers itself when an operator authenticates herself to the WFM through the application framework and requests a work unit. Each operator has one or more roles which govern the type of work that she may do, so, for example, an operator with the “transcriber” role may be assigned transcription work units but will not be assigned name normalization work units.



Design

To build a transcription application (TA) for this project, the Lee Library project team described their current workflow and added a wish list of features designed to make transcription more efficient. Basically the TA is a text editor with the ability to embed XML tags around specific text such as names, dates, places, etc. It is also very flexible so that simply by changing settings in its configuration file, new tags may be added to the TA. Although the Mormon diaries project is not using this feature, the TA may also be configured for single key (one operator transcribes the text), double key (two operators transcribe the text and the 2nd operator reconciles errors), or double key with blind reconcile (two operators independently key the text and a 3rd operator reconciles the differences between the two transcriptions) which typically gives character accuracy of 95%, 99.5%, and 99.95% respectively.

Because the name “Joseph” may in one place refer to “Joseph Smith” and in another place to “Joseph Campbell”, an authority manager (AM) for name normalization is also needed. With this application, an operator can review names tagged by the TA and approve or change them as necessary. Both the TA and AM have been written as Java Webstart applications for platform independence and for ease of distributing updates.

Initial design for iTranscribe was completed in July 2004 and a first version of the software was used in the library by student operators in August 2004. Defects have been fixed and features have been added periodically since then. Distribution of new software is simple because a Webstart application self-updates as necessary.

The automatic application for the Post Process state transforms the data collected by the WFM from the workflow applications associated with the preceding states into a digital object, TEI-Lite in this case, for the CONTENTdm target host system. Transformations are easily customizable for different target systems.

Designing a new workflow management process, such as we did for the Mormon diaries project, is an ideal time to carefully examine legacy workflow processes in order to streamline the process, reduce the amount of manual labor needed, and increase efficiency. So far little attention has given to these points, but since the software is still not complete, we may yet be able to give them the attention they deserve.

Conclusions

There are many other manuscript materials in repositories around the world that would serve as likely candidates for transcription and distribution. Aside from other diary collections there are letters and other correspondence, ledgers, political and business records – the list goes on and on.

Certainly distribution methods other than CONTENTdm will better meet the needs of different collections and users and we hope that through the approach of maintaining original images and properly marked up texts, flexibility of distribution will be maintained long-term.

As the variety of material being transcribed broadens, we also anticipate distributing the workload to expanding groups of interested parties, including family historians and genealogists or perhaps even high school teachers and students who could benefit greatly from “hands-on” experience with primary materials.