Vlasta Radan

COLORADO'S HISTORIC NEWSPAPER COLLECTION

In the fall of 1998, the Colorado State Library received a Library Services and Technology Act (LSTA) grant to develop an operational framework for the Colorado Digitalization Project (CDP). The ultimate goal of the project is to bring together all the different organizations entrusted with caring for the Colorado heritage and unite them in their effort to digitize their holdings and make them available thought the Internet. The experts from libraries, archives, historical societies, and museums were brought together to create the guidelines and handbooks for best practice in digitization, metadata input, management of the projects, and community outreach. Besides that, the CDP also served as the coordinator for some of the projects, like the Colorado's Historic Newspaper Collection or the Colorado Main Streets. Some of the projects, like Western Trails, are going beyond borders of Colorado and are designed to present the united heritage of geographical region, encompassing several states.

The Colorado's Historic Newspaper Collection is one of three projects executed under umbrella of the CDP. The main incentive for the digitalization of the newspapers was the long expressed need to find a more efficient way to research newspapers than rolling through the microfilms or leafing through the crumbing originals. Many of the original microfilms in use in Colorado libraries were created in the 1960s and were showing their age. The new digital technologies in combination with the Internet enabled the completely new ways of presenting, researching, and indexing the newspaper collections. Once digitized, the collections would be available to the wide range of remote users, many of whom are not in the position to travel to the view them in their original depositories.

The Newspaper Project was started in 2003 with $370,000 in combined funding grants received from LSTA and IMLS. The LSTA funds in the amount of $120,000, were used to buy hardware and software licenses. In 2005, the grants ended and the project now sustains itself through a combination of contributions, state funds, and foundation grants. Thanks to the careful planning and outreach programs, that made project widely known, the Newspaper project is able to continue the digitalization program and continue to add to the collection as funds become available (Appendix A). However, the reduction the grant money was not without pain even if the funding is generally satisfactory. The main problem is the uneven flow of contributions, which made it very difficult to manage an efficient workflow. Also, the personnel changes in partner companies create serous disruptions in the execution, cost, and timing of the process.

In 2006, the database created through the project includes 86 newspapers published in English, German, Spanish, and Swedish, between the years 1859 and 1928. The ultimate goal of the project is to digitalize all microfilm copies of newspapers owned by the Colorado Historical Society up to the year 1923—approximately 200 newspapers on over 2 million pages.

The collection could be searched through the union catalog or using the specially designed interface on http://www.coloradohistoricnewspapers.org. The CDP portal and web sites of participating institutions also provide links to that web site. Although the project would prefer to digitalize the complete runs of individual papers, the digitalization is limited to the microfilm reels ending with December 1922, i.e, material in the public domain.

The project decided to use duplicates of existing microfilms stored in the Colorado State Archives, scan them and than distill images into a web friendly format. The microfilms were duplicated in Colorado, in the State Microfilm Duplication Center, and then duplicates were sent to the OCLC Preservation facilities to scan the films. So created TIFF images were then uploaded to the FTP server where they were accessed by the Olive Software Inc. (http://www.olivesoftware.com/) which then used their Olive ActivePaper Platform to distill them into the XML format. The software is proprietary, but has the advantage that it is developed specifically for reformatting the newspaper materials into the XML format. It enables the full text search of the text but at the same time preserve the visual context of the published page.

When searched, the information could be viewed as individual paragraphs, articles, or printed pages. To prevent possible preservation and migration problems, the Newspaper project negotiated with the Olive Inc. that in the case the company goes out of business the Colorado State Library (which is the signatory for the license) would have access to the source code (Bishoff memo, 2003).

The digitalization processes – scanning and distilling – are completely automatic. While that provided for the speed and uniformity of the process, it also created the problems with newspaper issues that appeared out of the series on miscellaneous microfilms. In general, the project report noted that the initial proposal underestimated the time and amount of human labor that the execution of the project would require in spite of all the automation.

The CDP developed the Western States Digital Imaging Best Practices guides for digital images and scanning resolutions and sizes that should be followed by all engaged in digitizing projects. To facilitate the uniformity and minimize the costs of projects, CDP also organized five regional scan centers that provide hardware support and prevent the duplication of the resources (Bishoff, 2000; McGill, 2004). However, due to the specific technology used, the Newspaper project did not follow this particular requirement. All technology specific work on the project was done in outside institutions – OCLC facilities in Pennsylvania and Olive Software Inc. in Israel. It was initially planed that the whole work would be done by the OCLC, but due to the cost concerns, the distilling part of the process was moved to Israel. The final report of the project notice that this move created problems and delays because the Olive Inc. staff was not used to particular standards and formats required for this project.

After the distilling process is finished, the Olive Inc. sends DVDs with the TIFF and index files back to the Colorado. The Colorado State Library staff is responsible to upload files and update the interface. The Newspaper project uses the Colorado State Library storage area network services to the run site, provide a T1 line for Internet access as well as for the back-up and offsite storage for disaster recovery (Final Project Report, 2005, p. 5).

From the beginning, the Newspaper Project planned to integrate technology support with the staff and facilities of the Colorado State Library's Networking and Resource Sharing Unit (NRS). Their systems administrator and the support and collection librarian were involved in the Newspaper Project from the initial phase. The primary mission of the NRS unit was to "create, develop, maintain and provide support for information resources and resource sharing programs delivered over the Internet to libraries and Colorado resident and students." (Appendix A) and the maintaince of the Newspaper Collection would naturally blend in their workflow. Nevertheless, the plan is to hire an additional person who would be primarily responsible for the duties associated with the Newspaper Collection. The NRS unit is funded by a combination of state and federal funds. The fees for annual software maintence, server, and storage devices are part of the regular NRS budget.

The CDP multi-media-multi-institution projects, which desired to offer unique access to the wide array of the heritage collections, required an updated version of the Dublin Core (CDP Dublin Core Version 2.1). In order to enable search across different collections, the CDP decided to create "the model of distributed images and centralized metadata" (CDP, Storage Policy). In a model, the digital images are stored in the local servers owned by an organization that owns the original collection that was digitalized, while CDP maintains the centralized union catalog (*Heritage*) with metadata describing the images. The union catalog does not resolve the problem of the field specific vocabulary and indexing, or the lack of authority files.

The Newspaper project, in particular, encountered the problem of the "historic language" – when certain terms or expressions change during the time and full text searches therefore do not bring the desired results. The problem could be resolved by authority files or thesauruses, but that would be very costly. The Newspaper focus groups brought up this problem as one of the most important issues, but it was decided that the project in this phase does not have time and money to resolve this problem.

All questions related to the technical support or collection will be answered by the NRS team, as is the case with all other CSL programs. However, all questions related to genealogy and historical research will be redirected to the Colorado Historical Society.

The Newspaper Project hardware and software is stored at the remote server location of the Colorado State Library, the practice common to all projects managed by the NRS. The tapes on which the system data is backed-up were bought with state funds and the process of back-up and data migration is integrated with other NRS projects and is part of their regular cycle. The fees associated with offsite storage of the back-ups are also integrated in the regular NRS budget.

In review, the project showed more need for reconnaissance work before forming the initial work plan. During the execution of the project, it was found that there is no central, comprehensive electronic inventory for the microfilms. Many of the films were older and in worst physical condition than anticipated.

One of the grant proposal goals the Newspaper Project was to “empower teachers at all learning levels to use the resource by exploring pedagogical issues.” The project used CDP educational consultants to coordinate the efforts. The Newspaper project formed K-12 Educator and Faculty focus groups, which worked with the Newspaper Project team in testing the search interface and wording of the tutorials. The project also developed the series of the workshops that promoted use of the newspaper database in educational process. In various locations throughout Colorado—local libraries, historical societies, local newspaper offices--the Project also organized all together 23 presentations explaining value of the Collection in life long learning or genealogical research. Such systematic approach made the Colorado’s Newspaper Collection relevant to the wider Colorado public and ensured that the whole community would be interested in continuous funding and further development of the project.

Bibliography

This report is based on the project documents available on the Colorado Digitalization Project web site (http://www.cdpheritage.org/index.cfm):

The Colorado’s Historical Newspaper Collection Final report with appendixes (A-U) http://www.cdpheritage.org/collection/chncFinalReport.cfm

Liz Bishoff. October 23, 2003. Colorado to Create Colorado's Historic Newspaper Collection: A Statewide Model for Digitization. Project Report memo. http://www.cdpheritage.org/collection/documents/CHNC_LSTA_rpt_2003-10-23.pdf

Digital Image Storage Policy. July 2001. http://www.cdpheritage.org/cdp/documents/policy_imagestorage_2001.pdf

Articles cited

Bishoff, Liz. June 2000. Interoperability and Standards in a Museum/Library Collaborative: The Colorado Digitization Project. *First Monday*. 5(6). http://www.firstmonday.org/issues/issue5_6/bishoff/

McGill, Tami M. (September 2004). Rapid Implementation of a large-scare text digitization project: Colorado State University Libraries’ Experience. *Colorado Libraries*. 30(1). (from WilsonWeb)