Best Practice Poster: 
Reusing Legacy Metadata for Digital Projects: The Colorado Coal Project Collection

Michael Dulock
University of Colorado
Boulder, USA
michael.dulock@colorado.edu

Keywords: metadata; legacy metadata; digital libraries; digital collections; archives; Dublin Core; McBee cards; keyslot cards; edge-notched cards; metadata repurposing; hidden collections

1. Introduction

Libraries and other cultural institutions are increasingly focused on efforts to unearth hidden and unique collections. Yet the metadata describing these collections, when such exist, may not be in an immediately useable format. In some cases the metadata records may be as exceptional as the materials themselves. This poster describes research underway into how libraries can repurpose metadata in archaic formats using the Colorado Coal Project Collection1 slides as a case study.

Metadata in outdated formats, whether analog or digital, are a mixed blessing for metadata practitioners when creating digital collections. On the one hand, practitioners are happy to have pre-existing descriptive information to accompany their materials, eliminating the need to re-describe a collection or the items it contains. On the other hand, a lot of work may be required to convert that metadata into a form that can be used by their digital systems. Examples of legacy metadata include archival finding aids in typescript, catalog cards, handwritten inventories, out-of-date database software, and other more exotic formats. The metadata thus preserved can provide a wealth of information for users of a digital collection, but first the data must be moved from its old format into a newer, digital system. Various tools, such as database conversion software or OCR (optical character recognition) applications, can be used to convert metadata. But those tools are not fool-proof. A text captured using OCR may still require manual quality checking, since OCR software may not be able to correctly interpret the inconsistencies of typescript. Even metadata captured in a spreadsheet may not be immediately useable. Manual intervention is required to separate different values in cells that contain multiple data points, for instance.

2. Background

The Colorado Coal Project Collection documents the history of coal mining in the western United States, primarily focusing on Colorado in the early 20th century. The original project was conducted between 1974 and 1984 by Eric Margolis and Ron McMahan of the Institute of Behavioral Science at the University of Colorado Boulder. The two researchers documented the history, technology, and lives of coal miners in Colorado through photographs and interviews with miners, community members, and historians to discuss topics ranging from mining camp life and immigration to working conditions, labor unions, and strikes. The physical collection, housed at the University of Colorado Boulder Archives, comprises over one hundred video and audio files of interviews, scores of transcripts, and over four thousand slides depicting mining life.

The slides are accompanied by over four thousand McBee cards, a manual computing format that saw occasional use for recordkeeping in the mid-20th century (McCoy, 1965; Rabinow, 1958; 

1 http://libcudl.colorado.edu:8180/luna/servlet/UCBOULDERCB1~76~76
Smith & Schnall, 1980). These cards contain written notes as well as punches around the edge which indicate various features of the slides such as locations, dates, and technical details. Transferring this rich metadata from thousands of cards into a workable digital format was a challenge. The poster examines the process of transferring the metadata recorded on these arcane cards to a 21st century digital library collection, utilizing a combination of student labor, Metadata Services staff, MS Excel, and careful quality control.

3. Methodology

The first part of the metadata transfer process was capturing the metadata on the cards in an electronic format that could then be manipulated. The data was recorded in a consistent manner according to a classification key included with the cards. Each card was divided into sections: text in the interior of the card recorded the slide number, title, date, and description information, image quality, and restriction/rights notes; a series of numerically-coded holes (locations for punches) were arranged around the edge of the card. These, too, were divided into sections according to type: decades, structures, historical notes, “general” notes; states and regions; “general categories”; and technical notes. (See poster for a card and key images.) Each numbered pinhole was assigned a value on the key. Categories on the cards were mapped to metadata elements from the Dublin Core Metadata Element Set (DCMI, 2012). The Metadata Librarian built an Excel spreadsheet to capture the card metadata by category, which could then be crosswalked to Dublin Core (DC). The spreadsheet had one column for each category (slide number, decades, etc.), with each row representing a single card. Multiple data points would be entered in a single column but separated by a delimiter so the Metadata Librarian could later create one column for each entry. A key was added at the top of the spreadsheet indicating valid values for each category (text, 1-14, L0-L8, etc.). The spreadsheet would be filled out with data exactly as it appeared on the card, including numeric codes.

The Metadata Services Department hired three student workers to manually transfer the data. Each would be expected to record metadata from approximately 1,400 cards. The students were provided with written procedures as well as a visual job aid to make the transfer of data from card to spreadsheet as clear as possible (see poster). Having the students enter codes directly from the cards without translating them with the key served to reduce the labor time per card and eliminate mistranslation errors. In addition, Excel functionality could be used to isolate invalid data in individual columns based on the valid value ranges for some columns.

The Metadata Librarian checked the students’ output periodically throughout the project. Quality issues were minor and mostly typographical errors with number entry. The biggest hurdle was the handwritten text on the cards: in some cases handwriting was difficult to decipher, especially for proper names. Students were instructed to note entries that were difficult to decipher, so that the Metadata Librarian could examine the cards and do additional research as needed. A portion of the problem cards were completed by a paraprofessional from the Metadata Services Department after a student recorded the numeric coding.

Once the card metadata was captured, the Metadata Librarian split columns with multiple entries into individual columns. This resulted in multiple columns for several categories such as structures and technical notes. Once each column contained a single data point, another round of quality control was performed. The Metadata Librarian used conditional formatting to highlight invalid entries in each column. In some cases, a variety of invalid entries were searched for (e.g., letters and numbers outside of the valid range) and some spot checking was done against individual cards.

Following quality control, numeric codes were replaced by textual terms from the key column-by-column. Since each card might represent multiple slides, the Digitization Lab Manager de-duplicated entries on the spreadsheet by comparing it with the actual slides, indicating redundant slide numbers, or those for which we had no corresponding slide. The Metadata Librarian then further divided the document’s rows into one per slide, removing entries for missing or redundant
slides. The Metadata Librarian then crosswalked the spreadsheet data into the DC form and loaded it into the digital library software. The entire collection, including non-slide material, was processed and published in the CU Digital Library in time for the centenary of the Ludlow Massacre of 20 April, 1914, a watershed event in mining history and labor relations in the United States.

4. Conclusion

The Colorado Coal Project Collection, as it exists in the University of Colorado Boulder Archives, is a large, complex, and rich resource for researchers in mining and labor in the United States. Capturing and displaying the robust metadata that accompanied it proved an interesting and significant challenge, and served as a lesson in dealing with legacy metadata.

Acknowledgements

The author would like to acknowledge the hard work of everyone who contributed to this project, especially students Sarah Han, Rosalyn Wong, and Grace Zhong; library technicians Cynthia Hardey and Jane Zumwalt; Digitization Lab Manager Patrick Mulcrone; Digital Initiatives Librarian Holley Long; and original researcher and project sponsor Ron McMahan.

References