Automated Enhancement of Controlled Vocabularies: Upgrading Legacy Metadata in CONTENTdm

Andrew Weidner
University of Houston, USA
ajweidner@uh.edu

Annie Wu
University of Houston, USA
awu@uh.edu

Santi Thompson
University of Houston, USA
sathompson3@uh.edu

Abstract

To ensure robust, reliable, retrievable and sharable metadata, the University of Houston (UH) Libraries initiated a Metadata Upgrade Project in 2013 to systematically audit and refine the quality of the metadata in the University of Houston Digital Library (UHDL). Still in progress, the Metadata Upgrade Project has already produced significant discoverability improvements in the UHDL's legacy metadata and laid the foundation for future metadata production according to recognized standards. The final phase of the project includes aligning controlled vocabulary terms with appropriate authorities and adding and revising descriptive content in the UHDL. This is a time intensive process that requires careful evaluation and entry of name and subject authority terms. To improve efficiency and accuracy during the data entry process, the metadata librarian at the UH Libraries developed name and subject authority applications that automatically transform legacy controlled vocabulary terms into authorized forms. This project report provides an overview of the UH Libraries Metadata Upgrade Project, a discussion of how the UHDL's upgraded metadata improves discoverability of our collections, and an in-depth look at the custom tools that automate the authority alignment process in the CONTENTdm Project Client.

Keywords: metadata; controlled vocabularies; authority control; automation

1. Introduction

The University of Houston (UH) Libraries are committed to the dissemination and discoverability of our unique, historical collections. In the five years since the launch of the University of Houston Digital Library (UHDL), the repository has grown steadily and currently provides online access to more than 50,000 digital objects. While the UHDL serves as a platform for researchers to access the rare and unique materials in the UH Libraries holdings, the state of the legacy metadata in the digital library presented barriers to efficient use of the UHDL's digital objects. Incomplete and inconsistent legacy metadata restrict both discoverability and interoperability. To ensure robust, reliable, retrievable and sharable metadata, the UH Libraries initiated a Metadata Upgrade Project in 2013 to audit and refine the quality of the metadata in the UHDL.

The Metadata Upgrade Project team developed a three-phase strategy to systematically manage the metadata audit and upgrade process based on feedback and data analysis from focus group interviews, data inspection and benchmarking. Still in progress, the Metadata Upgrade Project has already produced significant discoverability improvements in the UHDL's legacy metadata. The third phase requires time intensive work on item level descriptive metadata revision including aligning controlled vocabulary terms with appropriate authorities. To improve efficiency and accuracy during the data entry process, the metadata librarian at the UH Libraries developed name and subject authority applications that automatically transform legacy controlled vocabulary terms into authorized forms.

2. Metadata Upgrade Methodology and Strategy

The Metadata Upgrade Project utilized several approaches to identify metadata issues and create strategies to improve the quality of metadata in the repository. To understand metadata

needs and address concerns that developed around legacy metadata, librarians conducted focus groups with UH Libraries stakeholders—including Special Collections, Web Services, and Liaison Services. External stakeholders were not included in the focus group interviews because of the complicated institutional review board (IRB) application requirements and the difficulty in identifying users. The project team also benchmarked current practices with similar digital libraries. These two activities demonstrated that controlled vocabularies in the UHDL had been applied inconsistently and inaccurately over time, most likely as a result of frequent changes in staff from project to project. Consequently, some items in the UHDL had rich descriptive connections with items in different digital collections while others had no terms to link them to similar materials. The Metadata Upgrade Project team concluded that the controlled vocabulary terms in the UHDL should be revised for accuracy, standardized to specific vocabulary lists, and mapped to appropriate Dublin Core elements (Thompson and Wu, 2013).

TABLE 1: Three Phases of the Metadata Upgrade Project

Project Phase	Tasks
Phase One	Stakeholder Interviews, Metadata Schema Development
Phase Two	Collection-level Metadata Editing, Metadata Dictionary
Phase Three	Item-level Metadata Editing

After collecting data regarding the issues with the legacy metadata in the UHDL, librarians developed key recommendations, a three-phase strategy for upgrading UHDL metadata (Table 1), and a new input standard to ensure that the quality of future metadata remains accurate and consistent over time. The first phase of the upgrade process focused on adding, revising, and standardizing descriptive and administrative fields. The second phase edited metadata at the collection level. Tasks performed in phase two included standardizing collection names for archival and digital collections as well as editing collection-level fields. The third phase focuses on adding and revising descriptive content in the digital library at the item level. To ensure that future UHDL metadata complies with the new standard, the Metadata Upgrade Project also produced a Metadata Dictionary which provides definitions, examples, and input rules for descriptive, administrative, technical, and preservation metadata fields (Thompson and Wu, 2013). An abridged version of the UHDL Metadata Dictionary (2014) is available online.

3. Automated Metadata Transformation

Addressing issues with controlled vocabulary terms is a key activity in the third phase, and the Metadata Upgrade Project staff spends a considerable amount of time reviewing existing terms, identifying more appropriate terms, and reconciling terms with the source vocabularies. In the early stages of phase three, the Metadata Upgrade Project staff experimented with exporting data from CONTENTdm and cleaning the data with OpenRefine. However, getting the cleaned data back into the system with a batch process proved a difficult task. The staff chose to work in the CONTENTdm Project Client for all phase three item-level editing and use OpenRefine for metadata analysis on new collections.

In order to speed up the editing process, the UH Libraries metadata librarian developed two applications that enable efficient transformation of legacy authority data within the CONTENTdm Project Client. Both applications are written in AutoHotkey (AHK), an open source scripting and macro language for the Windows operating system. In addition to a GUI that provides user feedback and menu functions, the core AHK scripts act as a glue language that connects the data in the Project Client with locally maintained vocabulary mapping files. Each AHK authority app gathers data recorded in the CONTENTdm Project Client and parses the tab-delimited authority files for matching terms. As of this writing, the tab-delimited files contain approximately 900 subject mappings and 3,000 name authority mappings. The apps automatically enter authorized terms in the Project Client and facilitate the addition of new terms to the local mapping files with input boxes and automatic Web browser searches. Most importantly, the apps

allow the Metadata Upgrade Project team to focus on the intellectual content of their authority work and let the computer take care of repetitive data entry tasks.

3.1 Subject Authority App

The decision to develop a subject authority app stems from the desire to ensure that the metadata for every object in the UHDL contains subject terms from a widely used controlled vocabulary. Legacy subject data in the UHDL includes terms from multiple vocabularies, and the subject app performs automated mapping from those vocabularies to authorized terms in the Library of Congress Subject Headings (LCSH). The UH Libraries are exploring opportunities for applying linked data technologies to the collections in the UHDL, and the subject app also facilitates harvesting of URIs from the Library of Congress Linked Data Service in preparation for that work.

```
CopyField:
    Send, {F2}
    Sleep, 50
    Send, ^a
    Sleep, 50
    Send, ^c
    Sleep, 50
    Send, {Tab}
Return
```

FIG. 1. AHK sub-routine for copying data and moving between Project Client fields.

The subject authority app processes one record at a time in the Project Client's spreadsheet view. When a metadata specialist triggers the subject app with the specified key combination, the app traverses one row and copies the data in each alternate subject authority field to the clipboard. In addition to LCSH, the UHDL uses four other subject vocabularies: Thesaurus for Graphic Materials (TGM), Art & Architecture Thesaurus (AAT), the Thesaurus for Use in College and University Archives (SAA), and a local UHDL vocabulary. To move between fields and copy the data, the app sends key presses to the Project Client, as if a human user were pressing keys on the keyboard. The sub-routine in Figure 1 sends the F2 key to activate the Project Client field for editing, Control + A (^a) to select all of the text, Control + C (^c) to copy the text to the clipboard, and the Tab key to move to the right one field. Brief pauses in between each keystroke (Sleep, 50) give the Project Client GUI time to process each command.

```
AAT natural disasters   Natural disasters   http://id.loc.gov/authorities/subjects/sh85090214   AJW 20140422
AAT hurricanes  Hurricanes  http://id.loc.gov/authorities/subjects/sh85063195   AJW 20140422
AAT boxcars Railroad trains http://id.loc.gov/authorities/subjects/sh85111077   AJW 20140422
AAT tracks (transit system elements)    Railroad tracks http://id.loc.gov/authorities/subjects/sh85111063   AJW 20140422
UHDL    Drifting/Damage Hurricane damage    http://id.loc.gov/authorities/subjects/sh2007001716 AJW 20140422
UHDL    Buildings/Streets   Buildings; Streets  http://id.loc.gov/authorities/subjects/sh85017769   AJW 20140422
AAT seawalls    Sea-walls   http://id.loc.gov/authorities/subjects/sh85119273   AJW 20140422
```

FIG. 2. Subject mapping entries in the local tab-delimited file.

After copying values in a field, the app parses the clipboard data and attempts to match each term against a tab-delimited mapping file stored on a local network drive (Figure 2). If no match is found for a given term, the app opens a Library of Congress Linked Data Service search for that term in a Web browser. After identifying an appropriate controlled term, a metadata specialist enters the authorized form and authority record URI in dialog boxes. The app automatically adds the term and URI to the local mapping file. When all of the alternative subject authority columns have been queried, the app returns to the LCSH column and inputs the authorized LCSH terms for that record (Figure 3) (Weidner, UHDL_SubjectTopical_CDM, 2014).

LCSH	TGM-1	AAT	SAA	Local
Natural disasters; Hurricanes; Sea-walls; Hurricane damage; Buildings; Streets		natural disasters; hurricanes; seawalls;		Drifting/Damage; Buildings/Streets;

FIG. 3. Subject values in the Project Client after mapping.

3.2 Name Authority App

The UHDL name authority app performs similar matching and mapping functions in a different direction. Instead of mapping values in multiple columns to a single vocabulary, the name app maps values in a single column to multiple vocabularies: Library of Congress Name Authority File (LCNAF), the Handbook of Texas (HOT), and a local UHDL name authority file (Figure 4). Much of the legacy name authority data in the UHDL is recorded in the LCNAF field, even though many of those names do not have records in the LCNAF vocabulary. This occurred as a result of the metadata schema work in phase two of the Metadata Upgrade Project when staff divided the UHDL's name fields (Creator, Subject.Name, etc.) into multiple vocabularies instead of one general field. In an effort to produce high quality, standardized data that is compatible with linked data principles, the name authority app automates the transfer of name data to the appropriate authority column in the CONTENTdm Project Client (Weidner, UHDL_Names_CDM, 2014).

```
Loop, parse, namelist, `n
{
    lcnafconfirmed := NameMap(lcnaf, A_LoopField, name, lcnafconfirmed)
    hotconfirmed := NameMap(hot, A_LoopField, name, hotconfirmed)
    localconfirmed := NameMap(uhdl, A_LoopField, name, localconfirmed)
}
```

FIG. 4. AHK loop passes each name to the NameMap function which returns an authorized form.

Monitoring accuracy during authority work is very important, and the Metadata Upgrade Project staff periodically review the name app's tab-delimited mapping file in OpenRefine to identify names mistakenly mapped to more than one form. Faceting on the authorized form column quickly reveals any problems with the data. As a quality control feature, the name authority app creates a report for each day and a log entry each time the name app is triggered (Figure 5). Using these reports, staff can backtrack to locate any records that must be corrected.

```
_____________________________________
2014-06-16 11:23 Early Tex Proj 4 AD

NAMES: Patterson, John

LCNAF:
HOT: Osterhout, John Patterson (1826-1903)
Local:

_____________________________________
2014-06-16 11:25 Early Tex Proj 4 AD

NAMES: Patton, Robert S., d. ca. 1857

LCNAF: Patton, Robert S., -approximately 1857
HOT:
Local:

_____________________________________
2014-06-16 11:26 Early Tex Proj 4 AD

NAMES: Perry, E. W.

LCNAF:
HOT:
Local: Perry, E. W.
```

FIG. 5. Name app report illustrating correct mappings to authorized forms.

3.3 Authority App Limitations

During the course of the authority work with the name and subject applications, the Metadata Upgrade Project team has identified a number of limitations. The apps can handle the bulk of the work, but there are edge cases that present interesting problems. In the case of the subject authorities, mappings to LCSH may change between collections because a single term in an alternate vocabulary can map to the multiple LCSH authorized terms. For example, the term "gutters" in an alternate vocabulary could map to "Roof gutters" or "Street gutters" in LCSH, depending on the context of the collection. This problem requires careful evaluation of a record each time the app is triggered and occasional editing of the tab-delimited subject mapping file.

In the case of the name authorities, there are many times when a name is present in both the LCNAF and HOT vocabularies. An update to the app provided the ability to harvest URIs from both vocabularies and record those connections in a separate file for future use. The app gives precedence to LCNAF for data entry purposes. As previously mentioned, the local tab-delimited name mapping file requires constant monitoring to ensure the accuracy of the authorized forms entered in the UHDL's metadata. Both AHK authority apps are short term solutions for the Metadata Upgrade Project and must eventually be supplanted by more robust controlled vocabulary management features in the UHDL's digital asset management system.

4. Benefits of Enhanced Metadata

There are numerous benefits to upgrading the legacy metadata in the UHDL. Integrating metadata best practices—including the consistent use of established controlled vocabularies—shaped the strategies and standards developed to address the issues identified during focus group interviews and benchmarking. These best practices will improve how users connect with UHDL content. In particular, standardized vocabulary terms consistently applied improve recall during faceted browsing, reducing the likelihood of orphaned records. Implementing best practices also ensures that UHDL metadata is fully interoperable with harvesting protocols, such as OAI-PMH, thereby providing another potential discovery layer to our content and opening up possibilities for collaboration with larger projects.

Aligning controlled vocabulary terms with recognized authorities and harvesting authority record URIs also lays the foundation for publishing UHDL collections as linked data with rich semantic markup. A first step might be to enrich subject terms and names with an owl:sameAs link, populated by the URI gathered during the Metadata Upgrade Project, that points to the unambiguous definition in the source vocabulary (W3C, 2004). Finally, with the creation of a more robust metadata dictionary, UHDL metadata creators now have a standard to guide future projects (Thompson and Wu, 2013).

5. Conclusion

While it is crucial to employ standards and best practices for quality control during the creation of a repository's metadata, metadata must be constantly maintained to reflect changes in the data model, end-user interface configuration, and system transitions. The lack of batch processing and limited authority control features in our digital asset management system creates barriers in our metadata editing workflow. The rapidly growing volume and complexity of formats in our digital library also presents challenges for our data quality management work. The utilization of scripting and automation in our metadata revision process has assisted us greatly in overcoming these barriers and challenges. The subject and name authority applications described in this paper have simplified our workflow and helped to improve consistency and accuracy in our data.

Metadata is at the functional core of our digital system. High quality metadata not only enhances the user experience in our digital library, but also enables the scalability and interoperability of our data. To ensure high quality metadata, it is important for metadata professionals to leverage traditional skills and new technologies to address the complex issues involved in metadata creation and maintenance. Applying traditional cataloging skills during

descriptive metadata creation and enhancing data with applications for automated analysis and transformation—such as data mining, name and subject heading mapping, and batch processing—will improve the quality of the metadata in our repository and the efficiency with which it is created. The UH Libraries will continue to explore and experiment with new approaches to describing our digital objects and, with the metadata upgrade work outlined in this paper, we are laying the groundwork for the migration of our data to a more expansive semantic environment.

References

Art & Architecture Thesaurus. (2014). http://www.getty.edu/research/tools/vocabularies/aat/index.html/. Accessed July 26, 2014.

AutoHotkey. (2014). http://www.autohotkey.com/. Accessed May 29, 2014.

Handbook of Texas. (2014). http://www.tshaonline.org/handbook/. Accessed July 26, 2014.

Library of Congress Linked Data Service. (2014). http://id.loc.gov/. Accessed July 26, 2014.

Library of Congress Name Authority File. (2014). http://id.loc.gov/authorities/names.html/. Accessed July 26, 2014.

Library of Congress Subject Headings. (2014). http://id.loc.gov/authorities/subjects.html/. Accessed July 26, 2014.

OpenRefine. (2014). http://openrefine.org/. Accessed August 10, 2014.

Thesaurus for Graphic Materials. (2014). http://www.loc.gov/pictures/collection/tgm/. Accessed July 26, 2014.

Thesaurus for Use in College and University Archives. (2014). http://www.archivists.org/publications/epubs/thesaurus.asp/. Accessed July 26, 2014.

Thompson, Santi and Annie Wu. (2013). Metadata overhaul: upgrading metadata in the University of Houston Digital Library. Journal of Digital Media Management, 2(2): 137-147.

UHDL Metadata Dictionary. (2014). http://digital.lib.uh.edu/about/metadata/. Accessed August 7, 2014.

W3C. (2004). OWL Web Ontology Language Reference. Retrieved July 26, 2014 from http://www.w3.org/TR/owl-ref/#sameAs-def/.

Weidner, Andrew J. (2014). UHDL_Names_CDM. GitHub Repository. Retrieved May 29, 2014, from https://github.com/metaweidner/UHDL_Names_CDM/.

Weidner, Andrew J. (2014). UHDL_SubjectTopical_CDM. GitHub Repository. Retrieved May 29, 2014, from https://github.com/metaweidner/UHDL_SubjectTopical_CDM/.