Metadata Quality Control for Content Migration: The Metadata Migration Project at the University of Houston Libraries

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

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

Abstract

The decision to migrate digital objects from one digital asset management system to another creates an excellent opportunity to clean and standardize descriptive metadata. The processes involved in moving large amounts of data from one platform to another lend themselves to automated analysis and remediation of metadata problems. The University of Houston (UH) Libraries established a Digital Asset Management System (DAMS) Implementation Task Force in early 2014 to explore possibilities for implementing a more robust repository architecture for the UH Digital Library. During the digital asset management system testing process, the UH Libraries Metadata Services Coordinator developed a set of scripts to programmatically access the data in the UH Digital Library through the existing digital asset management system API, create reports that were used to identify and correct problems, and lay the foundation for publishing UH Digital Library metadata as linked data. This project report discusses the background for the DAMS Implementation Task Force's work and the metadata quality improvements that resulted from it as part of a new Metadata Migration Project.

Keywords: metadata migration; quality control; digital asset management; automation; controlled vocabularies; linked data

1. Introduction

Metadata quality is an often overlooked or neglected aspect of digital repository development. In the excitement of setting up a repository infrastructure, the focus typically points to the software and hardware that allow institutions to publish digital collections on the World Wide Web, such as scanners, cameras, servers and turn-key content management system software. In the absence of trained metadata staff, descriptive metadata creation becomes a secondary activity that must be done in order to get a collection online rather than an essential process that facilitates effective discovery of a repository's resources.

Over time, as a repository's content grows, repository managers may realize that the quality of their descriptive data has suffered in the absence of careful attention to detail and consistent application of recognized standards. This is especially true when an institution explores opportunities for migrating data from one digital asset management system to another, as data analysis begins and decisions must be made regarding metadata transformations. This project report describes how the University of Houston (UH) Libraries leveraged the decision to test new digital asset management system software to analyze metadata in the UH Digital Library (UHDL), correct the problems it found, and prepare the UHDL descriptive metadata for publication as linked data.

2. Digital Asset Management System Evaluation

Since the launch of the UHDL in 2009, the UH Libraries have made thousands of rare and unique items available online using CONTENTdm, a proprietary digital asset management system owned and maintained by OCLC. While CONTENTdm helped the UH Libraries establish digital collections, the system has its limitations. The UH Libraries' digital initiatives have expanded, and the UHDL requires a more dynamic and flexible digital asset management system

that can manage larger amounts of materials in a variety of formats. The new digital repository infrastructure must also accommodate creative workflows and allow for the configuration of additional functionalities such as digital exhibits, data mining, cross-linking, geospatial visualization, and multi-media presentation. In addition, a system designed with linked data in mind will allow the UH Libraries to publish its digital collections as linked open data within the larger semantic web environment.

The *University of Houston Libraries Strategic Directions, 2013-2016* set forth a mandate to "work assiduously to expand our unique and comprehensive collections that support curricula and spotlight research. We will pursue seamless access and expand digital collections to increase national recognition" (p. 7). To fulfill the UH Libraries' mission and the mandate of the strategic directions, a Digital Asset Management System (DAMS) Implementation Task Force was created to explore, evaluate, test, and recommend a more robust DAMS that can provide multiple levels of access to the UH Libraries unique collections at a larger scale. The collaborative task force consists of representatives from four library departments: Metadata & Digitization Services (MDS), Web Services, Digital Repository Services, and Special Collections.

3. Metadata Upgrade Project

Concurrent with the work of the DAMS Implementation Task Force, the Metadata Unit in MDS wrapped up a two year project to normalize and standardize the legacy descriptive metadata in the UHDL. The Metadata Upgrade Project was initiated in 2013 to systematically analyze the descriptive metadata in the UHDL, standardize Dublin Core field usage across the UHDL's collections, and correct metadata content errors (Weidner et al., 2014). The analysis (Phase 1) and standardization (Phase 2) phases of the project produced a Metadata Dictionary (2014) input standard that guided the remediation work undertaken in the third phase, as well as metadata creation for new UHDL collections.

During the remediation phase (Phase 3) of the Metadata Upgrade Project, the Metadata Unit staff edited descriptive metadata for 54 collections comprising more than 9,100 digital objects. The Metadata Upgrade staff followed a workflow outlined at the beginning of the project. Tasks varied from collection to collection, depending on the state of the original metadata. Many tasks were accomplished through automation, such as aligning subject terms with controlled vocabularies (Weidner et al., 2014). After the Metadata Upgrade Project's metadata remediation phase was complete, the Metadata Unit staff conducted an audit of the tasks outlined in the project plan. Anomalies were noted, along with tasks that fell outside of the original project scope, for a subsequent undertaking to further refine the descriptive metadata in the UHDL.

4. Systems Testing

In late 2014, the DAMS Implementation Task Force began testing two systems as part of its charge to select a new repository architecture for the UHDL: DSpace 4 and Fedora 3. Web Services installed both systems in a development environment, and test collections from the UHDL were selected for ingestion into both systems. Rather than start from scratch with the original files and spreadsheet metadata, the Metadata Services Coordinator developed a set of Ruby scripts that access the data in the UHDL through the CONTENTdm API. These "cdmeta" scripts harvest image, audio, and video files as well as descriptive data and transform the descriptive data into DSpace Dublin Core and Fedora FOXML metadata (Weidner, 2015). Using these scripts, metadata and files for the test collections were quickly produced in the ingest formats required by DSpace and Fedora.

Recognizing the potential for applying the same technique to the Metadata Upgrade Project's authority control work, the Metadata Services Coordinator re-wrote the systems testing scripts as a Ruby library for object oriented access to the CONTENTdm API and created scripts that harvest names and subject terms in the UHDL. The "cdmeta_reports" scripts collate the harvested vocabulary data in plain text reports that list which objects are described by each term (Weidner,

2015). A second set of scripts filters the harvested lists of names and subject terms for unique values and writes those values to text files for each controlled vocabulary. Preliminary inspection of the vocabulary harvest files revealed common authority control problems, such as misspelled terms and multiple versions of the same name. Further inspection revealed terms that do not exist in the vocabulary to which they were assigned in the UHDL. Between the issues identified in the Metadata Upgrade Project audit and the controlled vocabulary terms harvest during systems testing, MDS recognized the need for a new project to prepare the UHDL's descriptive data for systems migration. As shown in Table 1, the work completed during the Metadata Upgrade Project and Systems Testing set the stage for the Metadata Migration Project that is currently underway.

TABLE 1. UH Libraries Metadata Projects Goals

Project	Goals
Metadata Upgrade	Standardize Metadata Schema Establish Input Standard Implement Controlled Vocabularies Correct Mistakes
Systems Testing	Develop Tools for Data Extraction Develop Tools for Analyzing Repository Data
Metadata Migration	Align Data with Controlled Vocabularies Prepare for Data Migration Prepare for Linked Data

5. Metadata Migration Project

The Metadata Migration Project at the UH Libraries began in early 2015 after the completion of the Metadata Upgrade Project. Expected to last until mid-2017, the project aims to build on the workflows and tools developed during the Metadata Upgrade Project and DAMS Implementation Task Force systems testing to further refine the UHDL's descriptive metadata in preparation for migration to a new digital repository architecture. The project will consist of iterative cycles of analysis and remediation to align controlled vocabulary terms with recognized authorities and prepare for linked data. After the new repository architecture has been implemented, the Metadata Migration Project will be complete, and all UHDL content will be migrated to the new system (Figure 1).

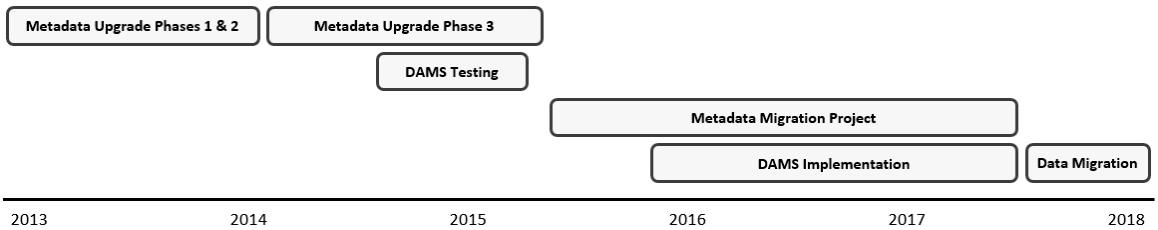

FIG. 1. UH Libraries Metadata Projects Timeline

5.1. Metadata Analysis

Using the cdmeta reports scripts described in Section 4, the Metadata Services Coordinator compiled lists of all the subject terms and names in use in the UHDL and further separated all of the unique values into vocabulary specific lists. This began a stage of metadata analysis that required staff time and the development of additional tools to partially automate the verification of controlled vocabulary terms. Verification of subject terms and names followed a two-step

process designed to identify problems with the values in use in the UHDL and gather URIs for valid terms that will be used in future linked data applications. A different employee performed each step so as to guarantee the authoritative nature of the UHDL's confirmed authority links.

The first step's primary goal was to gather URIs from the source vocabulary for authorized terms in use in the UHDL. To accomplish this task quickly and accurately, the Metadata Services Coordinator wrote an AutoHotkey (2015) application that automated repetitive tasks and allowed Metadata Unit staff to focus on verifying content. The application parses a controlled vocabulary list and displays each unverified term in a dialog box (Figure 2). At the same time, the application opens a search for the term in the vocabulary's online user interface in a web browser. The user can position the dialog box in a convenient location on the screen so as to quickly verify whether or not the UHDL term matches the term in the source vocabulary. If a match is found, the user clicks the Yes button and the application instructs the user to navigate to the linked data web page for that term. In the case of the Art and Architecture Thesaurus (AAT), that page is the Semantic View for Getty's Linked Open Data Vocabularies (Getty Vocabularies, 2015), as shown in Figure 3.

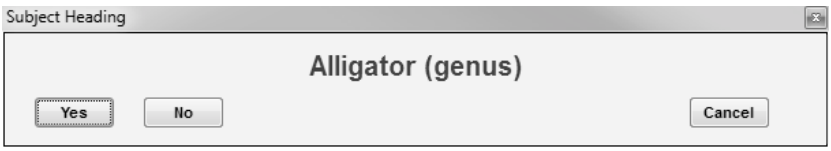

FIG. 2. Authority Verification Application Dialog Box

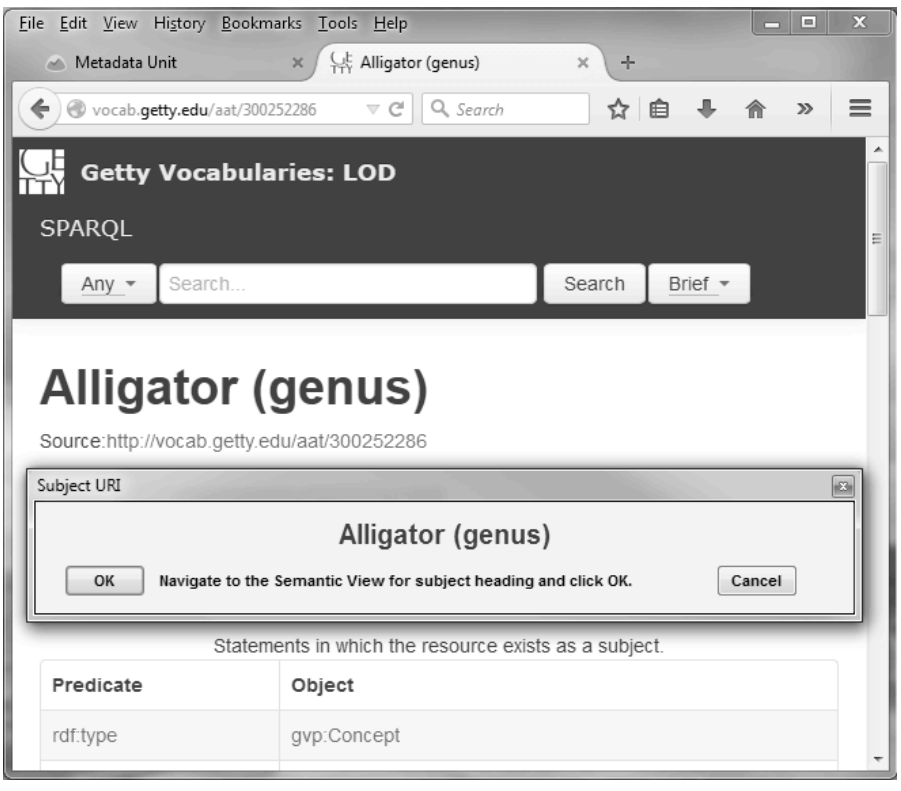

FIG. 3. Authority Verification Application & AAT Semantic View

After the user clicks OK, the application automatically harvests the subject heading URI, closes the tab in the web browser, and begins the process again for the next unverified term.

Verified terms and their associated URIs are recorded in a tab delimited text file. If the user discovers a problem with the term in use in the UHDL, clicking No in the initial dialog opens a second dialog, shown in Figure 4, which provides radio button options for indicating what is wrong. Common problems include misspelled headings and headings that have less or more information than the authorized form. Problem terms are recorded in a separate text file for further analysis and remediation work described in the next section. The second step in the controlled vocabulary term verification process utilizes a similar AutoHotkey application that displays a term in a dialog box, opens the linked data web page for that term, and asks the user to verify that the term in the dialog box matches the term on the web page. Any problems discovered during this stage are recorded in a separate text file, and the twice-verified tab delimited list of terms and their associated URIs are ready to be reformatted for use as linked data.

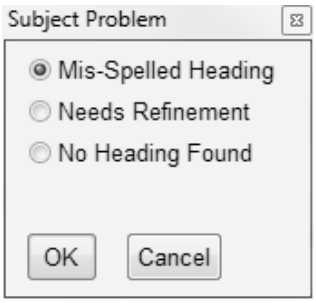

FIG. 4. Authority Verification Application Problems Dialog

5.2. Metadata Remediation

Despite the best efforts of the Metadata Upgrade Project, the programmatic harvest and analysis of the controlled vocabulary terms in use in the UHDL revealed many problems remaining to be corrected. The problems ranged in difficulty from misspelled subject headings to headings assigned out of context. Some of the context problems occurred because of an automation application used during the Metadata Upgrade Project that only allowed for one mapping from an alternate subject vocabulary to LCSH (Weidner et al., 2014). Other cases were the result of inadequate training of staff in descriptive practice and lack of effective metadata quality control at various times since the creation of the UHDL in 2009.

In order to make the large and complex remediation process more manageable, the Metadata Services Coordinator wrote an AutoHotkey script that cross-references the list of problems compiled during the authority verification process with the list of all subject terms in use in the UHDL. The script creates a new tab delimited text file for each controlled vocabulary that lists the subject term error and URLs to each object in the UHDL with that term in its metadata record, as shown in Figure 5. When viewed in Notepad++ (2015), the URLs became clickable links that Metadata Unit staff used to quickly locate the objects that required attention. The Metadata Unit Staff added authorized terms and URIs to the same tab delimited file after correcting the errors within the CONTENTdm Project Client.

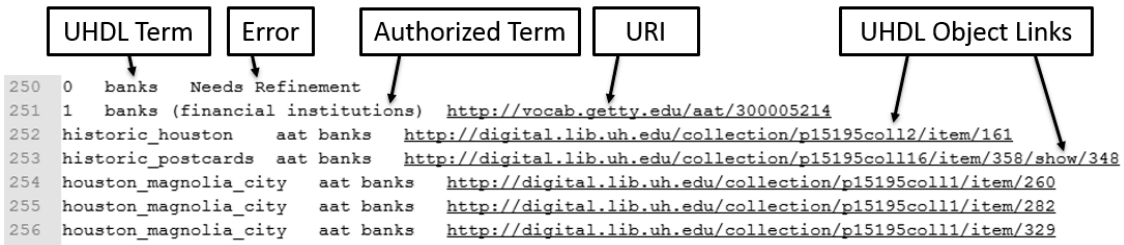

FIG. 5. Subject Term Errors with Authorized Forms and Links to UHDL Objects

For name authority reconciliation, the Metadata Unit leveraged a set of open source OpenRefine scripts that automatically harvest URIs from the Library of Congress Name Authority File (LCNAF) by querying the Virtual International Authority File (Carruthers, 2015). After separating the UHDL name lists into personal and corporate names, the OpenRefine scripts produced lists of matches with LCNAF URIs. The Metadata Services Coordinator developed an AutoHotkey script to divide all of the names into three categories: probable matches, questionable matches, and unmatched terms. Of the 1,223 unique names in the UHDL's descriptive metadata, the OpenRefine scripts found probable matches for 355 names, questionable matches for 347, and 521 names remained unmatched. Similar to the verification and remediation work for the UHDL's subject terms, AutoHotkey apps were developed to confirm linked data URIs and identify records that required metadata corrections in the name fields.

5.3. Linked Data and DAMS Implementation

An integral part of the Metadata Migration Project is preparing for the linked data environment. As previously mentioned, the Metadata Unit staff used a variety of applications to systematically harvest and verify URIs for authorized subject and name terms in a number of controlled vocabularies. Whenever possible, the process of recording the URI was automated to avoid copy and paste errors. This was accomplished with AutoHotkey by copying the text in the browser's address bar, as shown in the AutoHotkey function in Figure 6.

```
124 GetURI:
125     WinGetPos, xSubURI, ySubURI,,, Subject URI
126     winX := xSubURI
127     winY := ySubURI
128     Gui, 2:Destroy
129 
130     WinActivate, LC Linked Data Service
131     Sleep, 50
132     Send, ^l
133     Sleep, 50
134     Send, ^c
135     Sleep, 50
136     StringTrimRight, uri, Clipboard, 5
137     IfNotInString, uri, vocabulary/graphicMaterials
138     {
139         InvalidURI = 1
140     }
141 Return
```

FIG. 6. AutoHotkey Function to Harvest URI from Web Browser Address Bar

Eventually these links will enter the UHDL metadata to assert a relationship between the object and a subject term maintained in an external vocabulary. MDS is currently investigating the deployment of a vocabulary server to facilitate the consistent use of controlled vocabulary terms in the UHDL and throughout the UH Libraries (TemaTres, 2015). The UH Libraries will soon be implementing a new DAMS infrastructure based on Fedora 4 (2015), which conforms to the W3C recommendation for Linked Data Platforms (2015). Because of the work accomplished during the Metadata Migration Project, the UH Libraries will be in a good position to quickly publish our digital objects with links to external vocabularies when the migration to Fedora occurs.

6. Conclusion

The UH Libraries Metadata Migration Project is a natural continuation of the Metadata Upgrade Project. The improved quality of metadata, with URIs for controlled vocabulary terms, will prepare the UH Libraries for a smooth data migration to a new digital asset management system designed for the linked data environment. The implementation of the new system based on Fedora 4 will allow us to publish our digital collections as linked open data and open up new possibilities for effective use and re-use of the UH Libraries unique digital collections.

References

AutoHotkey. (2015). Retrieved April 7, 2015, from http://www.autohotkey.com.

Carruthers, Matt. (2015). LCNAF-Named-Entity-Reconciliation GitHub repository. Retrieved July 14, 2015, from https://github.com/mcarruthers/LCNAF-Named-Entity-Reconciliation.

Fedora Repository. (2015). Retrieved April 7, 2015, from http://fedorarepository.org.

Getty Vocabularies: Linked Open Data. (2015). Retrieved April 6, 2015, from http://vocab.getty.edu.

Metadata Dictionary. (2014). University of Houston Digital Library. Retrieved March 31, 2015, from http://digital.lib.uh.edu/about/metadata.

Notepad++. (2015). Retrieved April 7, 2015, from http://notepad-plus-plus.org.

TemaTres. (2015). Retrieved July 21, 2015, from http://www.vocabularyserver.com.

University of Houston Libraries Strategic Directions, 2013-2016. (2013). Retrieved March 31, 2015, from http://info.lib.uh.edu/sites/default/files/docs/strategic-directions/2013-2016-libraries-strategic-directions-final.pdf.

W3C. (2015). Linked Data Platform 1.0 Recommendation. Retrieved April 7, 2015, from http://www.w3.org/TR/ldp.

Weidner, Andrew, Annie Wu, and Santi Thompson. (2014). Automated Enhancement of Controlled Vocabularies: Upgrading Legacy Metadata in CONTENTdm. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2014, 167-172.

Weidner, Andrew. (2015). cdmeta GitHub repository. Retrieved April 1, 2015, from https://github.com/metaweidner/cdmeta.

Weidner, Andrew. (2015). cdmeta_reports GitHub repository. Retrieved April 1, 2015, from https://github.com/metaweidner/cdmeta_reports.