Use Existing Data First: Reconcile Metadata Before Creating New Controlled Vocabularies

Jeremy Myntti

Interim Head, Digital Library Services

J. Willard Marriott Library

University of Utah

295 S 1500 E

Salt Lake City, UT 84112

jeremy.myntti@utah.edu

Anna Neatrour

Metadata Librarian

J. Willard Marriott Library

University of Utah

295 S 1500 E

Salt Lake City, UT 84112

anna.neatrour@utah.edu

Introduction

The University of Utah has been building its digital library for over a decade. During this time, a wide variety of standards have been used for creating metadata. Uniform use of controlled vocabularies was not always a high priority in previous digital collection creation. Many creators, contributors, publishers, and subjects that should represent the same entity or concept are represented with different strings of data. Much of this data does not correspond to an existing controlled vocabulary. In order to find a solution to this problem, the library began an authority

control for digital collections project to standardize specific pieces of metadata and identify existing linked data-ready Uniform Resource Identifiers (URIs) that could be used to express the data.

This project started three years ago as a partnership between the University of Utah's J. Willard Marriott Library and Backstage Library Works (hereafter Backstage). In order to reconcile metadata values stored in the CONTENTdm XML metadata schema, Utah worked with Backstage to modify their authority control system, which has been used for many years to process MARC records into a system that could process XML metadata files.

In addition to cleaning up legacy metadata, a major outcome of this project was a report listing all of the URIs from the Library of Congress' (LC) Linked Data Service (http://id.loc.gov) and the Virtual International Authority File (VIAF) for the names and subjects that Backstage was able to successfully link to a matching access point. A second report listing access points without a corresponding URI was also created, helping metadata librarians identify names and subjects that will either need a URI linked from another controlled vocabulary or else a new URI will have to be minted in order to represent the entity or concept. During this process, small discrepancies (e.g. typos, missing elements of the name) were identified which caused many text strings to not match during Backstage's processing. After correcting these issues, we will be able to identify additional URIs from LC and VIAF using a couple of different methods.

We compared different processes for working with the named-entity recognition tool within OpenRefine to search these strings of data against VIAF in order to identify additional URIs that Backstage's processes may have missed. In addition, we tested named-entity matching within Alma, the University of Utah's integrated library system. Additional values were identified and added to our test collection.

The next step for this project will be to create a local controlled vocabulary with URIs to represent these entities and concepts. The system used to create these new controlled vocabularies will need to store data in an RDF format so when the data is published on the web, it will be accessible for others to reuse as linked open data.

Literature Review

Salo, (2010) stated that items being deposited in digital library repositories "tend to be disorganised, poorly described if described at all, and in formats poorly suited to long-term reuse. Even more unfortunately, researchers have become accustomed to the processes that produce these sloppy data, which makes them liable to resist changing those processes to improve data viability".

Using controlled vocabularies is important in digital library collections in order to provide standardized terms used for consistent description, searching, faceting, and interoperability. Making greater use of controlled vocabularies and linking to external sources of controlled vocabulary information becomes even more important when preparing data in digital collections for a linked data environment.

In Metadata Decisions for Digital Libraries: A Survey Report, Zeng, Lee, and Hayes (2009) identified that over 30% of those surveyed considered understanding the value of controlled vocabularies and planning how metadata records would be linked with authority records as a major concern when planning for metadata in digital libraries. Deciding which elements should use a controlled vocabulary or authority file was a major concern of over 66 percent of respondents and major concerns about data values centered on decisions to use existing

controlled vocabularies like the LC authorities or establishing local authority files and controlled vocabularies.

In a survey of the metadata practices used in digital libraries, Lopatin noted that nearly 90% of academic libraries use Library of Congress Subject Headings (LCSH) as the primary controlled vocabulary. While there were several other external controlled vocabularies used in academic institutions, locally created controlled vocabularies were among the top four most common (Lopatin, 2010). After comparing multiple controlled vocabularies for use in scientific metadata, White determined that while LCSH may be adequate for expressing topical subject headings, it is far from ideal as a controlled vocabulary for specific scientific terms. However, this study demonstrated that it is better to use a controlled vocabulary such as LCSH rather than rely on free-text keywords (White, 2013). Another study worked on finding ways to utilize the semantic modeling of LCSH in order to best express this type of controlled vocabulary within digital library metadata (Papadakis, Kyprianos, Mavropodi, & Stefanidakis, 2009).

The development of OpenRefine has opened up new possibilities for librarians to do greater work with authority control outside of the existing software they may already be using for display of their digital collections and accompanying metadata. In “Evaluating the success of vocabulary reconciliation for cultural heritage collections,” van Hooland, Verborgh, De Wilde, Hercher, Mannens, and Van de Walle demonstrated through a case study the issues involved in cleaning and reconciling the metadata in a large digital collection as the first step in leveraging linked data-enabled controlled vocabularies (van Hooland et al., 2013). In a blog post, “Archives Hub and VIAF Name Matching”, Jane Stephenson details issues with matching names with a VIAF reconciliation service, covering issues with matching when names have epithets, hyphenated names, life dates, and other common issues (Stevenson, 2013).

In order to discover how two different controlled vocabularies were related, Morshed and Sini used both a statistical approach and linguistic approach to match terms from the AGROVAC (a multilingual agricultural thesaurus) and CABI (a thesaurus created by a not-for-profit organization which provides information about agriculture and the environment) controlled vocabularies to discover their overlap. By using a process that compared relationships between terms, they were able to fully match 13% of the terms with many more potential matches based on different algorithms such as SMOADistance and LevesteinDistance (Morshed, & Sini, 2009).

Moine et al., (2014) discuss the steps used to create the controlled vocabulary for climate data in the METAFOR project. These steps are typical of any controlled vocabulary creation.

1. identify the relevant and discriminating information [...];
2. set an ensemble of appropriate terms (meaningful and non-ambiguous) to synthetically and faithfully express the information;
3. organize these terms hierarchically, with possible inter-dependencies;
4. attach a definition to each term;
5. identify allowed/possible values for each term.

In order for current cataloging practices to line up with linked data requirements, focus needs to shift from local cataloging silos to an environment where “decentralization, collaboration, localization, richness, and structure” provided by linked data are embraced (Seeman & Goddard, 2015 p 339). Several academic institutions have been conducting projects to find ways to utilize Resource Description Framework (RDF) data within their own repositories. These include projects such as integrating researcher data in VIVO (Ilik, 2015) to transforming existing metadata from a digital repository to RDF (Lampert & Southwick, 2013) (Southwick, 2015).

In a linked data environment, it is wise to have as many pieces of metadata belong to a specific controlled vocabulary with the data being represented by URIs as possible in order to increase "the likelihood of the [linked data] triple successfully interlinking with other relevant triples" (Southwick, 2015, p 7). According to Southwick, there are two main times when a new URI needs to be created to represent a string of data: for unique things owned by the local institution or for names that are not already controlled by a standards organization such as the Library of Congress.

Authority Control for Digital Collections with Backstage

In 2012, the University of Utah's J. Willard Marriott Library began exploring options to clean-up and reconcile specific metadata fields against the Library of Congress Name Authority File (NAF) and Library of Congress Subject Headings (LCSH). Authority control vendors have been doing this type of work in MARC records for many years, so we decided to start working with Backstage Library Works to see if they would be able to help us reach our goals of updating and standardizing our non-MARC metadata records. When we first approached Backstage, they were interested and excited to participate in this project since they have been successfully processing the authorities in the MARC21 standard and were looking for ways to expand their services. Processing authorities in non-MARC XML metadata schemas was a logical next step for them. The method for changing Backstage's processes to be able to handle XML files rather than only MARC21 records has been detailed in a previous study (Myntti & Cothran, 2013).

Utah currently uses CONTENTdm as its digital asset management system. While CONTENTdm uses an XML structure to store descriptive metadata for each item, the XML structure is unique to CONTENTdm and was therefore not as easy to work with as other standard XML formats. For instance, the CONTENTdm XML structure does not include any hierarchy to distinguish one record from another. To help Backstage's processes be able to distinguish between different

records, a simple hierarchy had to be created before processing the metadata. This hierarchy included adding a record tag (<record>...</record>) at the beginning and end of each unique record. While this helped to validate the records during the processing, the hierarchy had to be removed from the metadata prior to reloading it into CONTENTdm.

The basic workflow for having Backstage process the metadata was as follows:

1. Freeze any changes on the CONTENTdm server for the collection to be processed since any changes made would be overwritten when the corrected metadata files were re-indexed.
2. Make a copy of the CONTENTdm metadata file (desc.all) as a backup and send a copy to Backstage for processing. The desc.all file is the XML file that CONTENTdm uses to store all descriptive metadata for each collection. [Figure 1]

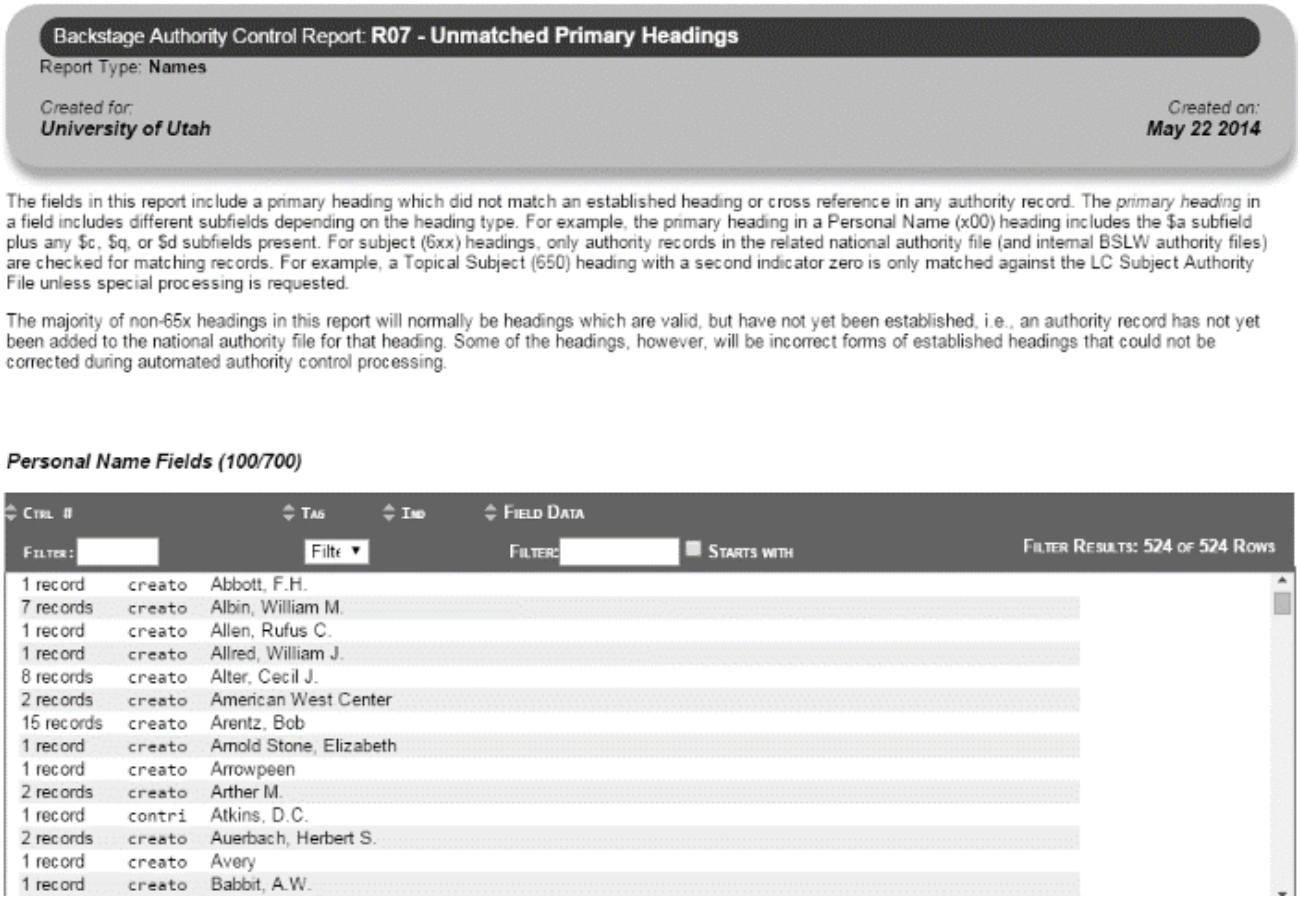
```
<title>The Social and Cultural Patterns of the Navajo Indians;</title>
<subjec>Indians of North America--Social life and customs; Liebler, H. Baxter (Harold Baxter), 1889-1982;
St. Christopher's Mission (Bluff, Utah); Religion; Missionaries; Missions;</subjec>
<covspa>Bluff (Utah);</covspa>
<keywor></keywor>
<tribe></tribe>
<band></band>
<creato>Liebler, H. Baxter;</creato>
<descri>The Social and Cultural Patterns of the Navajo Indians;</descri>
<publis>Digitized by: J. Willard Marriott Library, University of Utah</publis>
<contri></contri>
<date>1962;</date>
<dated>2010-03-23;</dated>
<type>Text;</type>
<format>application/pdf;</format>
<langua>eng;</langua>
<relati>This article is also a part of UTAH HISTORICAL QUATERLY VOL XXX (Utah State Historical Society -
Historic and Prehistoric Publications Collection);</relati>
<rights>Digital image copyright 2010, University of Utah. All rights reserved.;</rights>
<find>21292.jpg</find>
<dmcreated>2010-03-23</dmcreated>
<dmmodified>2014-02-17</dmmodified>
<dmrecord>18075</dmrecord>
```

3. Backstage then ran their authority control processes on the metadata file to update predetermined fields against specific controlled vocabularies.

4. After Backstage returned the metadata file to Utah, the desc.all file was replaced on the CONTENTdm server and the collection was indexed to reflect the changes that were completed by Backstage.

During the processing, Backstage was able to generate several reports that showed all of the changes that they made to the data as well as any data that did not match an access point in an LC authority record. Some of the reports that were the most helpful included:

1. *Near match report*. This report contained names or subjects that did not perfectly match an authorized access point or variant access point in NAF or LCSH. When an access point did not match an authority, Backstage ran the text string through another process to determine if there were near matches. By using a string-comparison algorithm, they were able to identify authority records where the authorized or variant access points were similar to the text string from the metadata record. This algorithm then assigned a probability percentage to the near match to help narrow down the list to those that are most likely to be a match.
2. *Unmatched headings report*. While some of these headings were also included in the near match report, there were many that were not. The unmatched heading report was useful in identifying metadata that was in the wrong field, such as a geographic place name or date in the creator field. With the unmatched headings report, a list was created with all access points that need to either be reconciled using another method (e.g. OpenRefine, manual review) or a new entry in a controlled vocabulary that needed to be created. [Figure 2]



Backstage Authority Control Report: **R07 - Unmatched Primary Headings**

Report Type: **Names**

Created for: ***University of Utah***

Created on: ***May 22 2014***

The fields in this report include a primary heading which did not match an established heading or cross reference in any authority record. The *primary heading* in a field includes different subfields depending on the heading type. For example, the primary heading in a Personal Name (x00) heading includes the $a subfield plus any $c, $q, or $d subfields present. For subject (6xx) headings, only authority records in the related national authority file (and internal BSLW authority files) are checked for matching records. For example, a Topical Subject (650) heading with a second indicator zero is only matched against the LC Subject Authority File unless special processing is requested.

The majority of non-65x headings in this report will normally be headings which are valid, but have not yet been established, i.e., an authority record has not yet been added to the national authority file for that heading. Some of the headings, however, will be incorrect forms of established headings that could not be corrected during automated authority control processing.

Personal Name Fields (100/700)

Ctrl #	Tag	Ind	Field Data
Filter:		Filte	Filter: Starts with — Filter Results: 524 of 524 Rows
1 record	creato		Abbott, F.H.
7 records	creato		Albin, William M.
1 record	creato		Allen, Rufus C.
1 record	creato		Allred, William J.
8 records	creato		Alter, Cecil J.
2 records	creato		American West Center
15 records	creato		Arentz, Bob
1 record	creato		Arnold Stone, Elizabeth
1 record	creato		Arrowpeen
2 records	creato		Arther M.
1 record	contri		Atkins, D.C.
2 records	creato		Auerbach, Herbert S.
1 record	creato		Avery
1 record	creato		Babbit, A.W.

3. *Updated headings report*. This report listed any changes that Backstage made to the metadata records, specifying the field that was changed, the old value, and the new value. This made reviewing the outcomes of the project easier in order to identify changes that may not be accurate. Many of the incorrect changes made during the automated processing were with undifferentiated personal names for local creators/contributors where they matched LC authority record when they should not have.
4. *Date change*. While this was not a change made to the records to reconcile data against an existing controlled vocabulary, Backstage's processes were able to review the format of dates in order to make sure they matched up with the ISO 8601 date/time format (e.g. YYYY-MM-DD). Any date that they changed was listed in this report for review as well as

dates that their program was not able to validate against the ISO standard.

In addition to cleaning up and standardizing the metadata, one major outcome of using Backstage to reconcile metadata fields against NAF and LCSH was that they were able to create a listing of all authorized access points in our metadata records and their corresponding URI from the LC Linked Data Service. These URI reports contained a long list of URIs that can be used in a linked data version of our metadata records in the future.

Results of Backstage processing

There have been multiple standards used to separate subjects from their subdivisions within the library's collections. This includes using a double dash, space double dash space, or an em dash from copying data from Microsoft Word. Backstage was able to standardize the usage to subdivision delimiters to only a double dash. This allows for a more consistent method of searching and displaying the data to library patrons.

Capitalization was another area where Backstage's processes were able to help standardize the metadata. Different standards of capitalization were used such as capitalize the first letter of each word, capitalize the first letter of the string, or have the entire subject string lowercase. Since the subjects were matched against LCSH, we were able to standardize all capitalization to LCSH standards.

Since this was an authority project, any access point that matched a variant access point (4XX) within an authority record was updated to the current authorized form (1XX) from the authority record. With many people working on metadata for digital collections over the years having varied knowledge about current cataloging practices, there were multiple instances where the variant access points were used rather than the authorized access point. This project was able

to successfully update all of those types of inconsistencies.

Challenges with Backstage processing

One of the major challenges of this project was taking an automated authority control system that has used MARC tags and subfields to identify the corresponding authority record and transforming it into a system that would be able to ignore MARC coding since data was stored in XML files. This lead to many instances where the wrong authority record matched a text string since the computer could not identify whether a subject was a personal name, topical term, geographic place name, etc. One example is the city name Provo. Since Backstage did not identify Provo as a place name, it was not able to limit the authority records searched to those that contained a 151 as the authorized access point. Instead, it treated this word as a topical subject and matched a cross reference in the “Provisional IRA” authority record. Another example is the subject “Cars.” The authority process matched this generic term with the authority record for “CARS 2002 (2002 : Paris, France).” There were not many instances of these types of errors, but they all had to be manually corrected after the processing was completed. One of the reports created during the processing showed all changes that were made to fields, so we were able to review those reports to discover these types of incorrect changes.

Another instance of this type of error occurred with names in a subject field. For many names that were expressed in a standard format such as [last name] [comma] [first name] [comma] [birth date] [hyphen] [death date], the process was able to figure out that the string was most likely a personal name. However, if a name was in direct order, had parenthetical information, or was a generic name lacking a date or other qualifier, then the matching algorithm was not able to automatically recognize that the string was a personal name rather than a topical subject. In order to find the most matches, some of the matching algorithms were loosened so that generic

names would have a better chance matching the correct authority record. For instance, the name “Bailey, Ron” could potentially match four different authority records, three with birth years and one with a middle initial, fuller form of the name, and a birth year. Backstage’s algorithm matched one of the names with only a birth year. In this case, the name was for a local person who did not have an LC authority record, and therefore should not have matched any authority. Other names incorrectly matched an LC authority record because they were for a local person and their name was an exact match for a different personal name in the NAF.

Another issue discovered was instances where there were multiple subject headings in a metadata record that all matched the same authority record (i.e. the authorized and variant access points were all included in the metadata record). When Backstage matched these text strings to the same authority record, the same authorized access point was entered into the metadata record multiple times. While this is not an issue that would hinder access to the item, it does make the metadata look messy with duplicated data. After discovering this issue, Backstage started looking into options to de-duplicate strings of data within a single field.

An issue that was quickly identified when reviewing the post-processing reports was data that had been entered in the wrong metadata field. There were instances where personal names were in a date field or topical subjects were entered in a spatial coverage field. These types of issues had to be reviewed and manually corrected after the processing had been completed.

Manual Clean-up

Since Backstage was able to generate several reports with details of the clean-up work completed as well as lists of names and subjects that they were not able to match against NAF and LCSH, there was a large volume of potential manual work that could be completed after processing. The library hired an intern who was pursuing an MLIS degree to review the project

outcomes and identify any problems or issues that would need manual corrections. The intern who reviewed these reports spent approximately 100 hours over the course of ten weeks reviewing the eighteen collections that had been processed.

Many of the changes that the intern was able to correct were cases where data had either been entered in a non-standard way (e.g. names in direct order rather than [last name] [comma] [first name]) or data that had been entered in an incorrect field (e.g. date in a spatial field). While the intern was manually correcting issues with names or subject headings, they made a list of all new access points and their corresponding URI from the LC Linked Data Service which the automated authority processing was not able to correct and identify. These types of manual corrections corresponding to a new match against LC were usually typo issues such as the wrong birth or death date or missing a middle initial or fuller form of the name.

Processing Statistics

While there were eighteen collections processed by Backstage during the initial phase of the project, one specific collection has been used for the remainder of this study: the Utah American Indian Digital Archive (UAIDA). In this collection, there were 7033 creator/contributor names and 98,931 subject headings used, with a majority of the pieces of data repeated in multiple records. For the creator and contributor names, 669 (9.5%) were updated to match the authorized access point from the NAF and 3685 (52.4%) matched and were linked to the correct name authority record. For subject headings, 21,072 (21.3%) headings were updated and 75,471 (76.3%) matched the corresponding LCSH authority record.

Vocabulary Reconciliation Process for Unmatched Names Comparison

Three approaches were explored for matching additional personal names beyond the ones that Backstage were able to identify. For this phase of the project we again used the Utah American

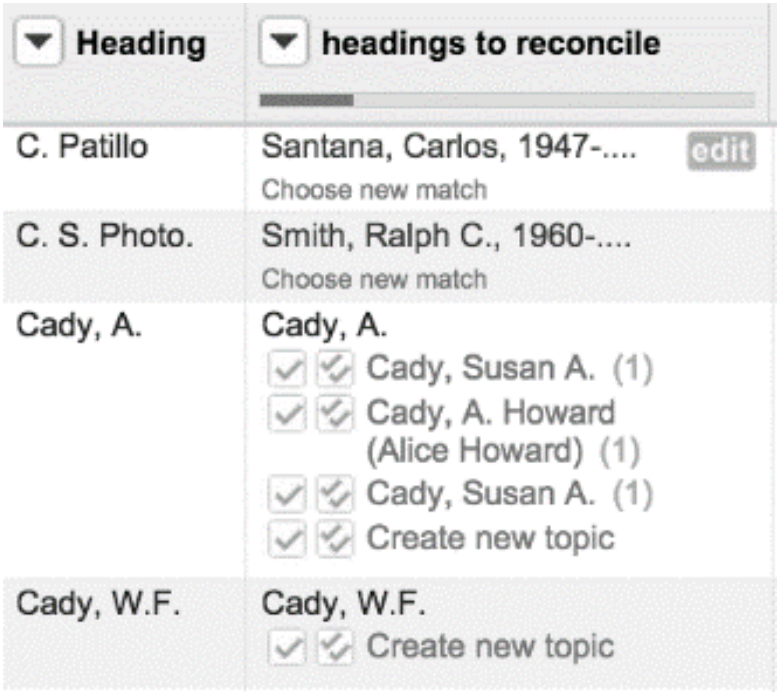
Indian Digital Archive (UAIDA), which has over 8,000 items including articles, books, maps, tribal documents, oral histories and photographs on Utah Native American tribes. The personal names in the collection contain a mix of regional names and more nationally known names that are more likely to be in existing name authority records. The metadata for the UAIDA collection was recently improved and cleaned-up in order to be conformant with the Mountain West Digital Library Application Profile, and has been harvested into the Digital Public Library of America. The collection at the time of this phase of the project had 529 unique unmatched names in the creator field after the previous authority control work from Backstage.

With the same set of unmatched UAIDA collection names, a metadata librarian tried two approaches to find additional matches using OpenRefine. First, the UAIDA unmatched names were run through the reconciliation service developed by Roderick D. M. Page, (Page, Roderic D. M., 2013), (Page, Roderic D. M., 2012).
In addition, the UAIDA names were reconciled using the process developed by Jennifer Wright and Matt Carruthers in Breaking the Bottleneck: Automating the Reconciliation of Named Entities to the Library of Congress Name Authority File, (Wright, Jennifer & Carruthers, Matt, 2015).

Both approaches required manual work for the metadata librarian after reconciliation to review matched and unmatched names. The new set of matched names was reviewed and flagged in OpenRefine if the match was false, and starred if the match was true. New rows in the OpenRefine projects were then generated with the URIs for the vocabulary items that met both conditions, so it was easy to compare the true and false matches with the original data set.

For the Wright and Carruthers' process, 81 records were matched, 132 were false matches, and 312 had no match. There were 262 undo/redo actions in the Google Refine Project when the metadata librarian completed manual review of the matches.

For the Page reconciliation service, 70 records were matched, 37 were false matches, and 449 had no match. The Page reconciliation represented a greater degree of manual work for the metadata librarian, because more possibilities were identified for each name. However, many of these possibilities turned out to be false matches or non-matches, as shown, where many possible matches ended up being discarded during the review process after reconciliation had run [Figure 3]. There were 424 undo/redo actions in the Google Refine Project for this process.



Both processes reconcile personal names against VIAF, however the greater degree of specificity offered by Wright and Carruthers with their process returning LC record links for matches created less manual work for a metadata librarian to review in this particular case. This had direct application in particular for the UAIDA collection since it contains more North American names that are most likely to be found in NAF. The reconciliation service developed

by Page provides more possibilities for matches, but by default many of the suggested matches represent the full range of name authorities available in VIAF, including sparse records from the German National Library and the National Library of France that ended up not contributing greatly to the matches found. Additional refinements could have been performed in OpenRefine for this process to discard non-US sources of authority control, but since the rate of initial successful matches was greater for the Wright and Carruthers process, it was the process identified as our preferred method moving forward. If we were repeating this process for a digital collection with personal names that were more international in scope, the reconciliation service from Page would likely be more advantageous.

Another approach tested was to use the authority control processes within ExLibris Alma, the library's integrated library system. In Alma, there is a process that runs on a daily basis to attempt to match specific MARC fields to the correct LC name or subject authority record. Since Alma uses MARC21 bibliographic records, a few records were created that contained all of the unmatched creator/contributor names and subject headings in the appropriate MARC fields. After these records were loaded into Alma and the "Preferred Term Correction" job ran, there was only one creator name and a dozen subject headings that Alma was able to accurately match and update to the authorized form of the access point. Since this match rate was so low, it was determined that this method of processing was not worth the time and effort since less than one percent of the fields were successfully matched to an existing access point in the LC controlled vocabularies.

Improving Metadata in Existing Collections

The Marriott Library has workflows and processes in place for updating metadata outside of CONTENTdm, such as with the Backstage authority project. This can be done through exporting, updating, and reloading the tab separated values file that contains the metadata for

the collection, or through scripting against the desc.all xml CONTENTdm file. Previous metadata update projects have involved inserting place names in the Western Soundscape Archive Collection (Neatrour, Morrow, Rockwell, & Witkowski, 2011), as well as adding Archival Resource Keys into existing collection items.

After trying three different approaches to reconciliation of personal names in the UAIDA collection, the process developed by Wright and Carruthers was the most efficient in terms of requiring less manual work and finding more matches for personal names. The next step for updating the UAIDA collection was to update the desc.all for the collection again with the updated values. The metadata librarian exported from CONTENTdm a spreadsheet containing the unique CONTENTdm number for each item and the contents of the Creator field. [Figure 4]

	A	B
1	creato	dmrecord
2	Atkins, John D. C. (John Dewitt Clinton), 1825-1908;	27233
3	Atkins, John D. C. (John Dewitt Clinton), 1825-1908;	27331
4	Atkins, John D. C. (John Dewitt Clinton), 1825-1908;	27452
5	Auerbach, Herbert S. (Herbert Samuel), 1882-1945;	18063
6	Auerbach, Herbert S. (Herbert Samuel), 1882-1945;	18064
7	Badlam, Alexander, 1809-1894;	48714
8	Bancroft, Hubert Howe, 1832-1918;	22233

The current contents of the creator field was reviewed against the updated personal name possibilities generated by both processes. In a few cases, personal names were already updated with the new values, or in other cases not found, due to the previous authority control work that was done in the collection. Additional false matches were identified and discarded. At the end of this process there were 72 values to be updated in the UAIDA collection. A few typos were also identified that were fixed in the updated desc.all file. The two column spreadsheet was prepared with the updated name values along with the CONTENTdm record number for each item, representing changes in 455 individual records. Of this final set, 405 records were updated with names match by both OpenRefine reconciliation processes, 38 were identified only by the Wright and Carruthers process, 5 had names matched only by the Page

reconciliation server, and 5 records were corrected for typos outside of any reconciliation process.

After creating the simple two column spreadsheet, the Application Programming department at the Marriott Library used a scripting process to iterate through the updated values and once more update the desc.all file for the UAIDA collection. The desc.all file was updated, and the collection was indexed again with the updated values. While we considered also adding the URI for each term to the collection to make it more Linked Data ready, CONTENTdm would not currently be able to leverage the URIs to do anything meaningful with that information, and would result in cluttered records. We decided to update the string values only, but are keeping the enhanced name values and URIs stored locally for use in the future when our digital library software is more Linked Data compatible.

The process developed by Wright and Carruthers will be repeated for the additional digital collections with unmatched names processed by Backstage. We are also exploring repeating this process for other controlled vocabulary terms, for example in matching place names against GeoNames.

Creating a Local Controlled Vocabulary

Some of the creator and contributor names which did not match an existing LC authority record were eligible for creating a new authority record through the Name Authority Cooperative Program (NACO), such as notable people from the Mountain West region. Many of the other names were not significant enough to contribute to a national authority file, so we need to find a way to develop and maintain a local authority file or controlled vocabulary. When creating this local controlled vocabulary, we want to make sure that we are using linked data standards in order to expose the data to the open web for reuse by others. Many of the names that we will

add to this proposed controlled vocabulary only have significance in our local region, and could potentially be used by other memory organizations in the Mountain West who have collections related to the westward migration in the nineteenth century.

The next step in this project will be to implement workflows and best practices for creating entries in local controlled vocabularies for any name or subject that is not currently in the LC NAF or LCSH.

CONTENTdm includes a controlled vocabulary feature that can be used to validate data within an existing field. This solution has not been useful for this project since the controlled vocabulary list must include every piece of data within a field and it is not possible to specify whether one string is authorized in an external controlled vocabulary or whether it is a unique access point. This would make it difficult, if not impossible, to separate out things that we have already matched against LC. This feature also does not provide a useful means for linking cross references or variant access points together and there is not an easy way to share controlled vocabularies across many existing collections. Another issue with controlled vocabularies in CONTENTdm is that each entry has a 128 character limit. While this limit would not be a problem for most terms, there are some entries that would need to be truncated in order to meet this requirement. Since the CONTENTdm controlled vocabulary feature applies to a single collection, not being able to uniformly refer to a single controlled vocabulary across many digital collections creates problems with interoperability as well as efficiency because a metadata librarian working on multiple similar collections would need to separately update each vocabulary attached to each collection.

After determining that the controlled vocabulary feature in CONTENTdm was not sufficient for the library's needs, other solutions are being explored. There are a few possible systems and

methods currently being analyzed for this phase of the project and the ideal solution for our library has not yet been identified. The solutions currently being examined include creating local authorities in the library's Integrated Library System (Ex Libris' Alma), Protégé, Terminology Management Platform, and OpenRefine connected to a linked data triple store. Since one of the goals of this project was to make our data more linked data ready, the ultimate goal will be to house this data within a linked data triple store in order to expose it on the open web. However, some of the other options may be more accessible to implement in order to begin creating and maintaining this data in the near future.

One step that would be easy to implement with this project would be to create local authority records within Alma, which would provide a means for updating data, creating variant access points, and providing additional information about the name or subject. This data would be stored in the MARC21 authority format, making the records similar to those maintained by LC. During the test within Alma to match the access points to existing LC authorities that was mentioned previously, we were able to devise a way to create a new local authority record for each of the unmatched access points by manipulating the data in MARCEdit and then loading the authority records into new authority files in Alma where they could be updated to include additional information about each name or subject and also maintained over time. These types of records could eventually be exported from Alma as MARCXML records which could then be transformed into RDF through an XSLT script.

Protégé, an open-source tool created at Stanford University, provides a means for creating linked data ontologies that can be shared and worked on collaboratively. This tool is provided as both web-based software hosted by Stanford or also a downloadable version which can be installed locally. This system provides a framework for developing local ontologies that use W3C standards for linked open data (Stanford Center for Biomedical Informatics Research, n.d.)

Terminology Management Platform from AthenaPlus is a new platform created for use within Europeana in order to create and maintain controlled vocabularies. This platform was released in January 2015 and provides a means for exposing custom controlled vocabularies using linked data best practices (Roche, Christophe & Damas, Luc, 2014).

The other option being explored is to convert the data that is currently contained in spreadsheets within OpenRefine to RDF using the RDF extension. This RDF data could then be housed and maintained in a triple store such as Mulgara. This method would be the ultimate goal for the project since the data would be created using linked data standards and could eventually be released on the open web for others to re-use.

Conclusion

Collection managers with long-standing digital library collection programs are often faced with the need to clean up and standardize legacy data in order to keep up-to-date with current standards as well as to prepare for upcoming standards such as Linked Data. Cleaning and reconciling data offers great benefits in interoperability and standardization for existing collections, along with the additional advantage of matching a greater percentage of terms to existing controlled vocabularies.

For a large digital library with a great number of legacy items with non-standard metadata, outsourcing authority control and generating as many matches to external controlled vocabularies through automatic means is an efficient and cost effective first step in preparing digital collections for future developments with Linked Data. The automated reports resulting from this process have contributed to additional authority control work and more extensive collection updates. Involving interns and metadata librarians to work with controlled vocabulary

items that could not be matched by Backstage Library Works ensured that human effort was saved only for those cases where automatic matches could not be easily made. By thoroughly exploring the processes involved in outsourcing an automated authority control for digital collections to a vendor as well as mapping out a post-processing workflow for terms that automated processes are not able to match, the J. Willard Marriott Library has been able to improve the quality of metadata in existing digital collections while also taking the first steps in preparing for a Linked Data ready digital library environment.

References

Ilik, V. (2015). Cataloger Makeover: Creating Non-MARC Name Authorities. *Cataloging & Classification Quarterly*, *53*(3-4), 382–398. http://doi.org/10.1080/01639374.2014.961626

Lampert, C. K., & Southwick, S. B. (2013). Leading to Linking: Introducing Linked Data to Academic Library Digital Collections. *Journal of Library Metadata*, *13*(2-3), 230–253. http://doi.org/10.1080/19386389.2013.826095

Lopatin, L. (2010). Metadata Practices in Academic and Non-Academic Libraries for Digital Projects: A Survey. *Cataloging & Classification Quarterly*, *48*(8), 716–742. http://doi.org/10.1080/01639374.2010.509029

Moine, M.-P., Valcke, S., Lawrence, B. N., Pascoe, C., Ford, R. W., Alias, A., … Guilyardi, E. (2014). Development and exploitation of a controlled vocabulary in support of climate modelling. *Geoscientific Model Development*, *7*(2), 479–493. http://doi.org/10.5194/gmd-7-479-2014

Morshed, Ahsan-ul, & Sini, Margherita. (2009). *Creating and Aligning Controlled Vocabularies* (p. 4). Agricultural Information Management Standards. Retrieved from

http://aims.fao.org/capacity-development/publications/creating-and-aligning-controlled-vocabularies

Myntti, J., & Cothran, N. (2013). Authority Control in a Digital Repository: Preparing for Linked Data. *Journal of Library Metadata*, *13*(2-3), 95–113. http://doi.org/10.1080/19386389.2013.826061

Neatrour, A., Morrow, A., Rockwell, K., & Witkowski, A. (2011). Automating the Production of Map Interfaces for Digital Collections Using Google APIs. *D-Lib Magazine*, *17*(9/10). http://doi.org/10.1045/september2011-neatrour

Page, Roderic D. M. (2012, June 2). Using Google Refine and taxonomic databases (EOL, NCBI, uBio, WORMS) to clean messy data. Retrieved from http://iphylo.blogspot.com/2012/02/using-google-refine-and-taxonomic.html

Page, Roderic D. M. (2013, April 17). Reconciling author names using Open Refine and VIAF. Retrieved from http://iphylo.blogspot.com/2013/04/reconciling-author-names-using-open.html

Papadakis, I., Kyprianos, K., Mavropodi, R., & Stefanidakis, M. (2009). Subject-based Information Retrieval within Digital Libraries Employing LCSHs. *D-Lib Magazine*, *15*(9/10). http://doi.org/10.1045/september2009-papadakis

Salo, Dorothea. (2010). Retooling Libraries for the Data Challenge. *Ariadne*, (64). Retrieved from http://www.ariadne.ac.uk/issue64/salo/

Seeman, D., & Goddard, L. (2015). Preparing the Way: Creating Future Compatible Cataloging Data in a Transitional Environment. *Cataloging & Classification Quarterly*, *53*(3-4), 331–340. http://doi.org/10.1080/01639374.2014.946573

Southwick, S. B. (2015). A Guide for Transforming Digital Collections Metadata into Linked Data Using Open Source Technologies. *Journal of Library Metadata*, *15*(1), 1–35. http://doi.org/10.1080/19386389.2015.1007009

Stanford Center for Biomedical Informatics Research. (n.d.). *Protégé*. Stanford University. Retrieved from http://protege.stanford.edu/

Stevenson, Jane. (2013, August 16). Archives Hub and VIAF Name Matching. Retrieved from http://blog.archiveshub.ac.uk/2013/08/16/hub-viaf-namematching/

Van Hooland, S., Verborgh, R., De Wilde, M., Hercher, J., Mannens, E., & Van de Walle, R. (2013). Evaluating the success of vocabulary reconciliation for cultural heritage collections. *Journal of the American Society for Information Science and Technology*, *64*(3), 464–479. http://doi.org/10.1002/asi.22763

White, H. (2013). Examining Scientific Vocabulary: Mapping Controlled Vocabularies with Free Text Keywords. *Cataloging & Classification Quarterly*, *51*(6), 655–674. http://doi.org/10.1080/01639374.2013.777004

Wright, Jennifer, & Carruthers, Matt. (2015). Breaking the Bottleneck: Automating the Reconciliation of Named Entities to the Library of Congress Name Authority File. In *ALCTS Metadata Interest Group ALA Midwinter 2015 Meeting*. Chicago, IL. Retrieved from https://github.com/mcarruthers/LCNAF-Named-Entity-Reconciliation