Authority Control in a Digital Repository: Preparing for Linked Data

Jeremy Myntti

Head of Cataloging and Metadata Services, J. Willard Marriott Library, University of Utah

Nate Cothran

Vice President, Automated Services, Backstage Library Works

Keywords: authority control, controlled vocabulary, digital library, Linked Data, metadata

Running title: Authority Control in a Digital Repository

Abstract

In an effort to identify an automated means for updating and standardizing metadata within a digital collection, the University of Utah's Marriott Library and Backstage Library Works partnered to develop a service that would replicate the benefits of an automated MARC21 authority control project for digital library metadata. This paper will discuss how the process to update MARC21 bibliographic records was adapted to update data encoded in XML. Future directions for this project will include taking a close look at how it can be used to link URIs with strings of data in order to prepare for a Linked Data environment.

Introduction

For many years, the MARC21 standard has benefited from automated authority control methods for updating and authorizing metadata fields. A major reason why these authority control processes have been so successful is because the records are coded in a consistent and standardized

format, namely the MARC21 exchange format. All major integrated library systems (ILS) are able to import the same MARC21 records and index them to make the metadata searchable.

This is not the case with digital library metadata. While many of the structures of digital library metadata are coded in the Extensible Mark-up Language (XML), field names, XML schemas and structure, and methods for metadata storage vary widely. Some digital asset management (DAM) systems store metadata in flat files with either a standard or proprietary XML structure. Other DAM systems import the XML metadata into a database structure that may change the formatting and field names used within the original XML file.

This paper describes a method currently being tested to update and standardize specific metadata fields in XML files through automated means while minimizing the amount of time and money that need to be spent on manual work. This process will be able to replicate some of the benefits of a MARC21 automated authority control project in metadata records encoded in XML. This includes matching specific pieces of metadata against the Library of Congress Name Authority File (NAF) and Library of Congress Subject Headings (LCSH). This process can help institutions prepare for Linked Data by identifying Universal Resource Identifiers (URIs) from the Library of Congress' Linked Data service (http://id.loc.gov) that can be used to represent the strings of data as well as provide alternate forms of the name and subject headings.

The premise of Linked Data is that information need only be updated once since the relevant information that Linked Data references resides in a single location. Typically, data information has been stored as strings which need to be periodically updated with the latest version of the textual strings. Linked Data obviates the need for replacing the data with other data by establishing a master record of the authorized data in a Web-accessible location. Any updates that are applied to the Linked Data repository are then promulgated to all data files which make use of the Linked Data information.

Institutions that transition to Linked Data would find their information continually updated to the most current, correct version without any active involvement on their parts.

Traditional vs Digital Authority Control

MARC21 records are comprised of data fields and values within those fields, which correspond to descriptive fields and access points within the bibliographic record. The MARC21 record structure is organized such that machines can parse the data and generate a human-readable display of the contents within each record.

Access points in the MARC21 bibliographic format have generally been assigned to 1XX, 6XX, 7XX, and 8XX fields. Within each of these ranges of fields, there are numerous other fields and subfields that can be used to describe an object. For example, within the 1XX range of fields, a cataloger can input the personal name (100), corporate name (110), conference or meeting name (111), or uniform title (130).

The main part of the access point is usually located in $a (subfield delimiter a). Other descriptive information, as necessary, may be encoded in subsequent subdivisions or subfields within the same access point. For instance:

600 $a Smith, John, $d 1947 Apr. 16-

651 $a United States $x History $y 20th century.

The 600 tells the cataloger/machine that this is a personal name subject heading; the $a denotes the surname, forename (as available); the $d represents the life dates associated with the name heading.

The 651 tells the cataloger/machine that this is a geographical subject heading; $a is the main geographic location; $x is the general subdivision; $y is the chronological subdivision. Each separate

subfield ($a, $d, $x, $y) has specific meanings depending on the field in which they are present. While the 600 field can contain $d to designate dates for the name, the same $d would be invalid if it existed in the 651 field.

Automated authority processing typically parses each heading, normalizes it, and then attempts to find a match against a nationally-recognized authority record database. The two headings above, after parsing and normalization have been applied, appear like:

600 $ SMITH, JOHN $ 1947 APR 16

651 $ UNITED STATES $ HISTORY $ 20TH CENTURY

The subfield marker (i.e., 'a' in '$a') is normalized out as some subfields may have been incorrectly coded. Note that the field itself is considered part of the data to be checked against the national database. This ensures that the proper usage of the potential matching authority record is considered ahead of other possible usages. For instance, a normalized heading with a 600 field tells the machine to consider the personal name usages of the authority primarily; a 651 field tells the machine to consider geographic subject usages of the authority primarily.

When the 600 field's data is parsed and normalized, it is then searched against the desired national authority database. In this case, there is an authority record for this heading (http://lccn.loc.gov/nb2001021004):

100 $a Smith, John, $d 1947 April 16-

400 $a Smith, John, $d 1947 Apr. 16-

Whereas bibliographic records can span 1XX, 6XX, 7XX, and 8XX fields, the authorized version of these headings is derived from the authority record's 1XX field. That 1XX field may be valid for usage as a name heading, subject heading, series heading, or title heading. It is not uncommon for authority

records to contain multiple, valid usages for the same heading. So a 1XX, 6XX, and 7XX heading in a bibliographic record may all be tied to the same established or authorized form of the name from the authority record 1XX field.

Authority records also contain variant headings, located in the 4XX fields. These variants represent alternative or different forms of headings which are not valid. Programmatically, these 4XX field variants point to the authority record's 1XX field, which again represents the current authorized form of the heading in question.

During authority processing, matches can be found against either the 1XX or 4XX in authority records. When a match is found against the 4XX variant, the system is programmed to look at the 1XX field and update the heading in the bibliographic record to reflect the information contained in the authority record 1XX field instead.

In the example above, the LC authorized heading (located in the 100 field of the authority record) spells out **Apr.** to **April**. The machine processing will find a match against the variant 400 field in the authority record:

400 $ SMITH, JOHN $ 1947 APR 16

This 400 field points to the authorized heading in the 100 field:

100 $ SMITH, JOHN $ 1947 APRIL 16

The un-normalized version of the authority heading is then used to exactly replace the corresponding bibliographic heading. The original, incorrect bibliographic heading (represented as the 400 field in the authority record) is now available to users searching for it, so long as the institution's ILS indexes those variant fields within the authority records.

XML schemas do not typically include field elements such as 1XX, 6XX, 7XX, or 8XX fields, so applying traditional authority processing to metadata in a DAM has not been successful to this point. While a MARC21 field lists the actual numeric field, XML data elements can be described using field names which may differ depending on the metadata standard or DAM system being used. All of the following field elements could potentially be describing any of the MARC21 fields listed in the 6XX range:

<subject>, <subjec-1>, <subject-title>, <topics>, <entry>, <s01>, <geog>, <genre>

Traditional authority processing habitually looks for tag numbers (e.g., 600), followed by the data for that particular field ($a Smith, John). Subsequent fields in MARC21 format designate separate data elements that sometimes need to be searched differently; for instance, searching a 600 field differently than a 651 field.

MARC21 fields, while sequential in raw data format, are considered separate and should be treated separately. Individual elements in XML files may list all of the separate field data in the same (singular) XML field element, similar to data on a catalog card, like:

<subject>Smith, John, 1932-; United States--History--20th century;</subject>

Not only are the individual data strings located in the same XML subject element, they are even parsed differently. Semicolons may be used to delineate the breaks between distinct headings. Double-dashes might represent breaks where subfield delimiters ($x, $y, $z) would typically occur in MARC21 format. Attempting to handle these two scenarios in XML field elements represents a radical change in how typical automated authority processing (with MARC21 formats) parses field element information.

With the help of the ILS, traditional authority processing pairs the matching authority record(s) with the updated bibliographic record(s). When a particular authority record has been updated at some future time and is redistributed, the ILS typically applies a global update to those bibliographic headings

affected by the change. Without access to the matching authorities, institutions can only rely on providing access based exclusively on the bibliographic headings.

XML files are typically self-contained; that is, all relevant information is confined within the same XML document. In fact, DAMs normally cannot link back to a separate authority file, even if converted to a similar XML format schema. This latter point represents a large hurdle to overcome in applying automated authority control to digital collections. Without the variants in 4XX fields within the authority record being available, patrons and users searching for these variant terms will not find the desired digital record.

Literature Review

The discoverability of resources in a digital library relies on accurate and consistent metadata just as has been the case with traditional bibliographic control within an ILS. Yasser (2011) reviewed the library literature for different problems that exist in metadata records for digital repositories around the world. One of the five major categories that Yasser identified was "inconsistent values, which addresses the inconsistent use of values to represent resource characteristics. […] [D]ifferent values associated with an element may equally represent a characteristic of the resource, but they may be different enough in recorded form to undermine system functionality" (p. 60). Yasser (2012) also said that "inconsistent value errors occur when element values are recorded in different forms such that a system's functionality such as collocation may be undermined" (p. 374).

In a study of metadata and cataloging professionals, Park and Tosaka (2010) found that 90% of digital repositories use some type of metadata quality control mechanism to ensure consistency in their metadata, which provides the biggest challenge in the metadata creation process. These quality control procedures are used to verify that data values used in different records are representing similar concepts. Structural consistency is also verified so that the same type of information is presented in the

same metadata elements. Park and Tosaka (2010) concluded that "metadata quality plays an essential role in building good digital collections. The core functions of bibliographic control in facilitating discovery, identification, selection, and use of digital resources needed by end users […] depend on it" (p. 710). Another study of cataloging and metadata librarian job descriptions conducted by Park, Lu, and Marion (2009) said that "metadata creation is a vital component for resource description and discovery in digital environment" (p. 854).

Controlled vocabularies are integral in metadata records for information retrieval. In a study of 127 libraries, Lopatin (2010) reported that 93% "of academic librarians and 90% of non-academic librarians reported the use of controlled vocabularies" (p. 727). Zeng, Lee, and Hayes (2009) state that "controlled vocabularies […] and rules are usually required by metadata standards and application profiles for the values associated with subjects, media formats, resource types, audience levels, and so on" (p. 182). They also reported that "library communities already have a long history of developing and employing controlled vocabularies and authority files. In addition to using these existing value encoding schemes, developing controlled vocabularies for a specific project seemed of major importance" (p. 183).

Baca (2003) summarized the need for using a controlled vocabulary:

> Why is the use of subject vocabularies and thesauri so important for end-user access? Because users often have an idea in mind of what they are looking for, but do not know exactly what it is called. Objects, artists, places, concepts, etc., can be called by more than one name, and names may change depending upon the time and place in which they are being used (p. 52).

Hillman, Marker, and Brady (2008) said that the basic goals of a controlled vocabulary are to "eliminate or reduce ambiguity; control the use of synonyms; establish formal relationships among

terms; and test and validate terms" (p. 17). They continued that by implementing controlled vocabularies, the balance between precision and recall are enhanced to improve information retrieval.

Several studies have been completed discussing the importance of authority control in a library's data. Gorman (2004) said that "bibliographic control is literally impossible without authority control" (p. 12). He also said that "random subject, name, title, and series denotations that are not subject to any kind of standardization—vocabulary control—will lead to progressively inchoate results as databases grow, and when a Dublin Core database is of a sufficient size, the results will be no more satisfactory than those using free-text searching on the Web" (p. 16-17).

Harper and Tillett (2007) found that "when we apply authority control, we are reminded how it brings precision to searches, how the syndetic structure of references enables navigation and provides explanations for variations and inconsistencies, how the controlled forms of names, titles, and subjects help collocate works in displays, how we can actually link to the authorized forms of names, titles, and subjects that are used in various tools, like directories, biographies, abstracting and indexing services, and so on" (p. 53). Harper and Tillett (2007) continued by saying the "benefits of authority control described above–search precision, more powerful navigation, collocation, and linking between various tools and resources–apply to metadata about the creators of resources as well as to subject access" (p. 56-57).

Park, Lu, and Marion (2009) discussed the need for authority control with regard to information sharing among libraries, demonstrating the need for consistency. Multiple studies such as Boydston and Leysen (2006) and Dragon (2009) have acknowledged that there are limited resources (e.g. time and money) in metadata departments for creating unique headings for creators and subjects. Because of these limited resources, there need to be mechanisms for automating parts of this process. Boydston and Leysen (2006) briefly discussed a couple of options for automating metadata creation. One example

is the Johns Hopkins University Lester S. Levy Collection of Sheet Music Automated Name Authority Control (ANAC) program which is able to identify matching Library of Congress name authority records that match headings in existing metadata records. The second example cited by Boydston and Leysen is Meta-Extract from Syracuse University's Center for Natural Language Processing. Boydston and Leysen acknowledge that neither of these two systems for automated metadata generation and validation is perfect, so they also require some amount of human intervention to guarantee accuracy. Dragon (2009) suggested that having an automated process to authorize names already residing in the LC authority file would help metadata catalogers identify headings for which they wouldn't need to do any additional research to establish a unique heading.

Birrell, Dunsire, and Menzies (2010) compared the use of name and subject authority control in a typical library catalog or ILS with the same type of authority control in an institutional repository (IR), and found the former is generally far more advanced than the latter. This lack of authority control in an IR can create some issues that are "clearly significant in terms of semantic interoperability and the sharing of metadata" (p. 387). Two of the major features of authority control that are problems with an IR include the lack of a controlled vocabulary in user-generated metadata and the inconsistencies created when terms within a controlled vocabulary change without updating the associated metadata records.

Dragon (2009) recognized that authority control in a digital collection is helpful for our users, but she also recognizes that there can be some difficulties in devising headings for local creators, locations, or subjects which don't have Library of Congress authority records. She continued that this "requires that the cataloger either determine the correct form of the heading according to complex rules, or else forgo assigning a specific heading for a named entity. Thus the very quality that makes these images valuable to the user […] increases their complexity from the perspective of the cataloger"

(p. 189). Salo (2007) found that “the naïve user of an institutional repository will swiftly find that the absence of name authority control inhibits retrieval of items by a single author” (p. 250).The use of authority control with a digital library can help prepare for a Linked Data environment. Cwiok (2005) proposed that “as metadata schemas become more sophisticated, the focus of implementers shifts to issues of interoperability, authority control, and semantics” (p. 105). Harper and Tillett (2007) have said that “the availability of library authority data in a more Web-friendly format has the potential to positively influence the organization of the broad spectrum of Web content already available” (p. 57).

Methodology

Although this proposed service can work with virtually any XML file that follows a consistent schema, the examples, issues, and solutions contained herein pertain to proprietary XML files used by OCLC’s CONTENTdm (CDM) DAM system. In other projects previously completed at the University of Utah’s J. Willard Marriott Library, workflows have been established for making automated changes to the raw XML metadata files within CDM. An example of this type of change has been discussed by Neatrour, et al. (2011). The basic outline for making automated changes to the CDM XML files include the following steps:

1) Stop updates to the collection and make it read-only
2) Make a copy of the desc.all metadata file for backup and for automated processing
3) Run the desc.all file through the automated authority control processing
4) Replace the desc.all file on the CDM server
5) Run the full collection index
6) Remove read-only status from the collection

As would be expected, the biggest challenge in completing this project was altering an existing authority control service provided by Backstage Library Works which typically processes MARC21 and

MARCXML bibliographic records into a system that could provide the same type of processing on different types of XML files. The overall structure of a heading between a MARC21 file and XML file is similar enough for there to be potential to control those headings within XML files.

For instance, a subject heading in a MARC21 bibliographic record looks like:

651 $a United States $x History $y 20th century.

When the subject heading is parsed and normalized according to Name Authority Cooperative Program (NACO) normalization rules, prior to attempting to match against LC authorities, it becomes:

$ UNITED STATES $ HISTORY $ 20TH CENTURY

The subfields are removed (though the subfield delimiters remain), along with any punctuation that may be present, and the string is capitalized.

When using the Mountain West Digital Library's Dublin Core Application Profile, the heading would appear as:

United States--History--20th century;

Since the individual subfields (i.e., between "United States" and "History") are unknown, the following assumption was made when attempting to convert this XML heading into a standardized format:

651 $a United States $x History $x 20th century

This heading is nearly identical to the MARC21 version, except that final **$y** is interpreted (incorrectly) as **$x**. But this is just the initial attempt to format the XML string. Once it is normalized, it becomes:

$ UNITED STATES $ HISTORY $ 20TH CENTURY

This now matches the original normalized version of the MARC21 heading. It also means that the heading can be checked against the LC authority database in order to properly update the heading, when necessary.

Another difference that had to be considered when processing XML metadata rather than a MARC21 record is that it is normal for XML headings to all be located on the same line, separated by a specific delimiter (in this case, a semicolon has been used to separate multiple pieces of data within the same field):

Navajo Indians--History; Religious ceremonies; Religion;

The delimiter between headings, represented by a semicolon, helps parse out each distinct heading in the XML element. Instead of one long, conjoined heading, it is broken up into three distinct headings:

Navajo Indians--History;

Religious ceremonies;

Religion;

During testing, we discovered that **Religious ceremonies** is not a valid subject heading, but has been replaced by **Rites and ceremonies**. So when the processing was completed, the revised (conjoined) XML field became:

Navajo Indians--History; Rites and ceremonies; Religion;

Earlier we described a major impediment for implementing automated authority control on digital collections to be the distinct lack of using the matching authority file in combination with the

updated XML file. Even though we updated the obsolete form of **Religious ceremonies** to **Rites and ceremonies**, it would be even more useful if **Religious ceremonies** was still available as a search term that could be indexed. This type of problem could also be taken care of if an authority file could be implemented within a DAM system rather than having to insert variant forms of headings within additional fields in the metadata records. For the purpose of this project, the variant forms of the headings were added in additional keyword fields.

In order to preserve these types of keywords that may be useful to users, we decided to insert the variants (i.e., 4XX fields) from the authority record into the updated version of the XML record.

Thus, a new XML keyword field would be added to the relevant XML record:

<subject>Navajo Indians--History; Rites and ceremonies; Religion;</subject>

<subject-keyword>Ceremonies; Ecclesiastical rites and ceremonies; Religious ceremonies; Religious rites; Rites of passage; Traditions;</subject-keyword>

This potential solution would satisfy the need to have the variant forms available for patrons to search since patrons cannot be expected to always use the latest versions of headings in order to find the work or item in question.

A stipulation with this solution means that libraries would then need to index this field in order for it to be useful. However, if the subject element contained tens or hundreds of variants, then the possibility for very large indexes of subject-keyword elements also exists. The larger the index, the more time it would take to index the updated XML file.

It was also possible that variant forms of the heading would contain duplicate information. For instance, **Marriott Library** would be considered a duplicate of **University of Utah. Marriott Library** and so **Marriott Library** could be eliminated from the list of keywords added to the XML file. In summary, if a

given keyword element is contained within a larger keyword element, then the shorter version could be eliminated from the keyword element to avoid duplication.

Throughout the processing, a number of issues were identified. These issues, discussed below, included altering an existing MARC21 authority process to accept XML elements, single word elements that may be incorrectly updated, missing XML container elements in the CDM XML file structure, and variant forms of headings that would be useful to patrons.

MARC21 Field vs Generic Field Element

A major difficulty was that automated matching was previously set up based on the field in question (i.e., 600, 651, etc.). Backstage employs a very strict set of matching criteria where the field in question is concerned. For instance, it is not acceptable for a 610 field to change (through authority matching) to a 651 field. The 610 field contains information for corporate name headings and the 651 field encompasses geographic subject headings. Matching on one and flipping to the other is frowned upon as these two headings convey very different kinds of information to users, and it increases the likelihood of finding improper matches.

Within subject elements in the XML file, it is perfectly natural for both of these types of headings to appear side-by-side (offset by semicolons). In an effort to simplify the initial search, each heading within the XML subject element was forced to match as if it started out as a 650 field (topical subject heading) since a single subject string with multiple subject headings may contain multiple types of headings, such as topical, personal name, or geographic headings. As part of the traditional authority processing, fields that do not find a match against an authority 150 field (which corresponds to a bibliographic 650 field) can then also be searched against an authority 151 field (geographic subject heading).

Due to the strict matching criteria mentioned above, the rules had to be relaxed so that all of the other potential MARC21 fields could be considered during the matching phase: 600, 610, 611, 630, 650, 651, 655 fields.

Single Word Elements

Spatial elements, where data is sometimes listed as a singular geographic city or location, posed problems for the automated processing. Lehi is a city in Utah, as well as in Arizona. However, there is no (geographical subject heading) authority match for just **Lehi**. Due to the relaxed matching criteria mentioned above, **Lehi** is normalized to **LEHI** and does find a match on a variant (410) for **LEHI**. Unfortunately, this points to the authorized form in the 110 which is listed as **Llano Estacado Heritage, Inc.** This poses a problem as this is not what is desired: to have a single spatial entry updated to a heading that is completely different from the original.

Container Elements

CDM XML files do not typically include container elements. Container elements help offset records within an XML file, like:

<record> data </record>

Machine parsing the XML file tells the program that <record> begins the record and </record> ends the record. A new record would then begin with <record>. Within each set of <record> … </record> container elements, there are other elements describing the digital file. It is also not necessary for container elements to be referenced the same way, just so long as it remains consistent throughout the XML file. These container elements are similar to the MARC21 leader and directory, letting the computer know where one record ends and the next begins.

Within CDM, however, all records will begin with the same field such as:

<title>

…

<title>

…

In this example, the <title> element represents both the start and the end of a record within the XML file. As XML files can be set up according to different metadata schemes, it also means that elements can have different field names depending on the XML schema. One metadata schema may use <title> while another may use <title-proper>, and so on.

Compounding this issue is that CDM XML files must be explicit. That is, there can be no deviation in the naming of the XML elements or the order in which they appear within the XML file. This also implies that CDM XML files are not dynamic, so they cannot grow in the number of elements inserted into the file through outside processing.

If new elements are expected to be populated and generated through external processing (such as automated authority control), then those fields must be added to the metadata template for the collection prior to exporting the XML file to the vendor. Empty XML elements are fine, but the new structure needs to be present before the export is performed so that the eventual re-import will have the same fields in the same order to be indexed.

In order to get around the lack of container elements within the XML files, a generic container element (i.e., <record> … </record>) was temporarily added to each record. In the following example, we see two distinct records in the XML file, both of which begin with <title>. As the container elements are missing, we added them in a pre-processing step:

```
<record>
        <title>
        …
</record>
<record>
        <title>
        …
</record>
```

Now the program will be able to separate the processing for each record. It will also update that particular record with the proper authorized matches (when they exist), as well as add in the corresponding keyword elements where applicable. Of course there are other elements within each record, but for the sake of brevity they are not listed here.

In post-processing, the container elements that had previously been added are removed. Again, this is to ensure that the same elements that were exported from CDM are imported in the same order, with no deviation between the two files, other than some elements contain updated metadata.

Variants

As XML files within a DAM system cannot easily make use of external authority files, the issue arose as to how to still make that information useful. The obvious answer was to include variants in the same XML file. We have already addressed the need to add new elements to the XML file, and this step is where those newly created elements would then be filled in.

Before processing, we identified the exact elements that would need authority processing. For instance, within an XML file, these field elements may be:

- <creator>
- <contributor>
- <subject>
- <topic>
- <spatial>

Depending on which elements are desired for inclusion in the matching process, empty keyword elements corresponding to each desired element needed to be added such as:

- <creator-keyword>
- <contributor-keyword>
- <subject-keyword>
- <topic-keyword>
- <spatial-keyword>

Thus, when the <creator> field is updated with the authorized heading, the corresponding variant forms of headings (if any) are then added to the <creator-keyword> element. When users or patrons perform a search through the digital collection for either the authorized or the variant versions, they are directed to the proper digital file. In effect, this is similar to combining the MARC21 bibliographic and authority records into large metadata records. This solution is temporary until an authority file could be implemented within the DAM system being used for the test.

Outcomes

In the pilot project, which was run through the process discussed in the methodology, one-third of the records from the University of Utah's Institutional Repository (USpace) were altered or updated. With over 6,300 records processed, there were 658 creator names (0.03%) and nearly 3,000 subject

headings (18.4%) that were updated. In addition to the names and subjects that were changed, it was also found that 6,250 names (28.0%) and 5,300 subjects (33.0%) already matched the authorized terms from the Library of Congress Name Authority File and Subject Headings.

While there were many fields that matched the authority file, there were also many names and subjects that were not able to match or be updated through this automated process. Approximately two-thirds of headings (16,000 names and 10,000 subjects) were not able to exactly match an existing authority record from the Library of Congress. Due to the unique nature of digital collections, this can be expected since the Library of Congress authority files would not have all of the locally created headings for a particular institution's collections. In addition to the local nature of some of the content, it was found that many headings were not formed consistently according to Library of Congress or NACO standards, stemming from insufficient funding for training and quality control.

Even within the field elements in the XML file for a single collection, there can be many variations for how the data has been formatted. The most egregious issue would be missing semicolons, which are used to delineate individual headings according to the standards used by the Mountain West Digital Library contributing institutions. An example of this would be:

Smith, John, 1932- United States--History--20th century;

In this example, there is a missing semicolon between **1932-** and **United**. Unfortunately, there is no easy way to programmatically discern if a semicolon is missing on purpose from a heading. Reporting out such a heading on an unmatched headings report would at least call attention to this for further review by the participating institution.

However, other improperly formed heading issues could be more readily addressed and corrected:

- (proper) United States--History--20th century;
- (improper) United States-History-20th century;
- (improper) United States—History—20th century;

In both of the latter examples above, instead of a double-dash '--', the program encounters a single dash '-' and long hyphen '—'. In these cases, cleanup routines were performed on the headings prior to normalization and parsing. This ensures that, even in the event of no update to the original heading (i.e., no matching authority record exists), the heading is still properly and consistently formatted when returned.

Similarly, improperly formed dates may occur throughout the data (according to the W3C Date Time profile of ISO 8601):

- (proper) 1980; 1981; 1982; 1983;
- (improper) 1980-3;
- (proper) 2013-05-26;
- (improper) May 26, 2013;
- (improper) 20130526;
- (improper) 05-26-2013;

Through sampling, we discovered that some XML files contained improper optical character recognition (OCR) data. An example of improper OCR that would cause a problem with this automated processing is the inclusion of field element punctuation: angle-brackets or '<', '>'.

Part of the parsing algorithm employed for properly identifying each element within an XML record takes the angle-brackets into consideration. So an element that begins with '<' and ends with '>' is set apart as being its own distinct field element (e.g., <title>).

With some description or transcription OCR field elements, an errant angle bracket might throw off the automated processing. In order to work around this issue, we enclosed the OCR field element within computer data elements (CDATA), which are unparsed character data in an XML document. The XML parser will not parse any data within the CDATA block, allowing the OCR text to pass through unchanged. The CDATA block markers (applied in pre-processing) are then removed in post-processing. The limitation with this solution is that if there is other information contained within the CDATA block that could potentially be updated through this authority control process, it would be completely ignored.

A number of issues identified in this project can also be addressed through proper reporting. Some possible reports which are similar to reports generated in an ILS that could help identify other issues that could be corrected include:

- List of names or subjects changed, based on cross references in an authority record
- List of names or subjects not matched to a Library of Congress authority record
- Errors, typos, and inconsistencies identified
- Dates not standardized to a specific format
- Number of URIs identified from http://id.loc.gov
- Possible names to submit through NACO

A statistical summary could helpfully address sections where the institution is interested in numbers of elements affected, or updated. Then unmatched names could be addressed simply through an unmatched headings report, or in more depth through a near-match headings report.

In particular, Backstage utilizes a Levenshtein-Distance (LD) algorithm when determining potential near-matches to report back out to the institution. LD calculates the distance between two separate strings. This distance is determined by the number of characters one string must change in

order to match the other string. For instance, if one string is **Untied States--History--20th century** and the other string is **United States--History--20th century**, then LD determines the number of changes necessary to bring "Untied" to match "United". The more characters that match between the two strings, the greater the confidence value (expressed as a percentage point within the report). The above example would generate a confidence value of 94.4%. However, if we were working with a shorter version such as **Untied States** and **United States**, the LD algorithm generates a lower confidence value (84.6%) since there are fewer overall characters that match between the two strings. Percentages based on the LD are calculated and those higher than 75% are listed on a report for further review. This helps the institution determine whether there are typos or other anomalies within the unmatched headings that may be easily updated through a find and replace update.

Future Plans

As more libraries start to look towards providing access to their content through Linked Data, it will become increasingly important to have consistent and standard headings that correspond to existing authority files such as those provided by the Library of Congress.

This project will help prepare for a Linked Data environment by standardizing existing headings to be consistent with the Library of Congress Name and Subject authority data. Another way that this project will help prepare for Linked Data is by identifying an existing URI that can be used to represent the heading and which can also be used to retrieve additional, useful information about the heading. In the current DAM system being used for this test, it is not easy to use the data represented by the URIs to provide our patrons with additional context about the terms. However, these URIs can be stored in a separate database in order to periodically retrieve the current authorized form of the heading so that the text string metadata stored in the metadata records can be updated without requiring manual work.

This replicates some of the features of maintaining separate bibliographic and authority databases as is the common standard in the MARC21 format.

A tool is currently in development that will be able to take a simple database containing the URIs identified in this project, the authorized form of the heading as of the last date that the URI was looked up, and the digital collection(s) containing that specific piece of metadata. With this database, the tool will be able to look up the URIs to find the current authorized form of the name or subject heading, identifying pieces of metadata in different collections that need to be updated. Once a heading has been identified for updating, the tool will automatically change all of the variant forms of the heading to be consistent with the current authorized form in the Library of Congress authority records.

Conclusion

Any type of manual updating to records takes a large amount of time and money. In order to improve efficiencies, methods can be devised to automate the matching of certain metadata fields with their corresponding controlled vocabularies. The methods discussed in this paper include a way to replicate the benefits of a typical MARC21 automated authority control project in digital library metadata encoded in a variety of standards.

Linked Data has the potential to change the way library data is structured in the near future. In order to prepare for this change, libraries need to make sure that their data is consistent and standardized so that it has the highest potential for linking to existing controlled vocabularies that are available as Linked Data. This project has demonstrated that it is possible to complete major updates to records in order to bring them in line with the authorized terms in commonly used controlled vocabularies without a large amount of manual work that costs both time and money.

References

Baca, M. (2003). Practical issues in applying metadata schemas and controlled vocabularies to cultural heritage information. *Cataloging & Classification Quarterly*, 36(3-4), 47-55.

Birrell, D., Dunsire, G., & Menzies, K. (2010). Match point: Duplication and the scholarly record: The online catalogue and repository interoperability study (OCRIS), and its findings on duplication and authority control in OPACs and IRs. *Cataloging & Classification Quarterly*, 48(5), 377-402.

Boydston, J. M. K., & Leysen, J. M. (2006). Observations on the catalogers' role in descriptive metadata creation in academic libraries. *Cataloging & Classification Quarterly*, 43(2), 3-17.

Cwiok, J. (2005). The defining element: A discussion of the creator element within metadata schemas. *Cataloging & Classification Quarterly*, 40(3-4), 103-133.

Dragon, P. (2009). Name authority control in local digitization projects and the Eastern North Carolina postcard collection. *Library Resources & Technical Services*, 53(3), 185-196.

Gorman, M. (2004). Authority control in the context of bibliographic control in the electronic environment. *Cataloging & Classification Quarterly*, 38(3-4), 11-22.

Harper, C. A., & Tillett, B. B. (2007). Library of Congress controlled vocabularies and their application to the Semantic Web. *Cataloging & Classification Quarterly*, 43(3-4), 47-68.

Hillmann, D. I., Marker, R., & Brady, C. (2008). Metadata standards and applications. *The Serials Librarian*, 54(1-2), 7-21.

Lopatin, L. (2010). Metadata practices in academic and non-academic libraries for digital projects: A survey. *Cataloging & Classification Quarterly*, 48(8), 716-742.

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).

Park, J.-R., Lu, C., & Marion, L. (2009). Cataloging professionals in the digital environment: A content analysis of job descriptions. *Journal of the American Society for Information Science and Technology*, 60(4), 844-857.

Park, J.-R., & Tosaka, Y. (2010). Metadata quality control in digital repositories and collections: Criteria, semantics, and mechanisms. *Cataloging & Classification Quarterly*, 48(8), 696-715.

Salo, D. (2007). Name authority control in institutional repositories. *Cataloging & Classification Quarterly*, 47(3-4), 249-261.

Yasser, C. M. (2011). An analysis of problems in metadata records. *Journal of Library Metadata*, 11(2), 51-62.

Yasser, C. M. (2012). An experimental study of metadata training effectiveness on errors in metadata records. *Journal of Library Metadata*, 12(4), 372-395.

Zeng, M. L., Lee, J., & Hayes, A. F. (2009). Metadata decisions for digital libraries: A survey report. *Journal of Library Metadata*, 9(3-4), 173-193.