Notes on Operations

Supporting Name Authority Control in XML Metadata: A Practical Approach at the University of Tennessee

By Marielle Veve

While many different endeavors to support name authority control in Extensible Markup Language (XML) metadata have been explored, none have been accepted as a best practice. For this reason, libraries continue to experiment with the schema, tool, or process that best suits their local authority control needs in XML. This paper discusses current endeavors to support name authority control in XML for digitized collections and demonstrates an innovative manual solution developed and implemented by the University of Tennessee Libraries to achieve this goal. Even though this method for authority control in XML metadata still relies on manual efforts, it effectively reduces time and research work by efficiently setting priorities, identifying critical descriptive areas in the digital transcriptions, and identifying the most appropriate biographical resources to consult. The effectiveness of this approach in improving the rest of the metadata production workflow is evaluated and presented.

Soon after starting digitization projects, many libraries and other institutions often find that keeping track of name access points in the Extensible Markup Language (XML) is a huge challenge, regardless of the XML schema used. This is particularly the case in many types of digitized objects such as manuscripts, music, and other types of special collections where the number of personal names is exponentially more than the number of items digitized; the names are dispersed all over the digitized transcriptions; and information about these names is ambiguous, vague, and incomplete. However, no matter how difficult keeping track of name access points in digitized materials is, it is necessary in order to keep digitized objects retrievable. Access points not only help in the retrieval process of documents, but also help keep materials by the same creators or about the same subjects together.

To keep a successful track of name access points in XML documents, libraries have been experimenting with many different endeavors to find an effective way to achieve this goal. So far the efforts created to support name authority control in XML metadata consist of (1) using XML schemas to encode authority data; (2) endeavors for shared, cooperative, national, and international XML name databases; (3) manual and automated conversion tools from Machine-Readable Cataloging (MARC) to XML; and (4) automated generation of authority control through especially designed systems. The problem with most of these endeavors is that they only address the issue of how to encode name access points utilizing XML authority schema; they do not address the issue of how to extract or harvest these names directly from the XML records and transform them into useful access points. The few endeavors that have tried, such as the systems for automated generation of authority control, have only been successful in extracting names from XML records but not in turning them into

Marielle Veve (mveve@utk.edu) is Assistant Professor and Catalog and Metadata Librarian, University of Tennessee, Knoxville.

The author wishes to thank her University of Tennessee colleague Marie Garrett for helping edit the first draft of this paper and Peggy Johnson for her advice and valuable suggestions on the manuscript.

Submitted May 13, 2008; tentatively accepted and returned to author for revision June 26, 2008; revision submitted July 19, 2008, and accepted for publication.

reliable access points. This is because their name matching processes fail. For this reason these endeavors will always depend on human intervention to work properly. In addition to not being completely reliable, many of these endeavors are costly and labor intensive, not to mention that most of them only display newly created access points locally.

The method introduced in this paper to support name authority control in XML metadata addresses the issue of extracting or harvesting names directly from XML documents manually and turning them into useful access points. In contrast to the previously mentioned methods, this method is effective, relatively simple, cost efficient, and has the ability to display the new access points at the national level by still using the richest authority file available—the Library of Congress's (LC) authority file (LCAF, http://authorities.loc.gov). This method consists of a simple manual approach to extract and create name access points that effectively reduces time and research efforts by efficiently setting priorities, identifying critical descriptive areas in the digital transcriptions, and identifying the most appropriate biographical resources to consult. When using this method, libraries will not have to go through the work of encoding authority data into XML schemas or translating authority data from one schema to another. Neither will they have to worry about hiring a programmer to build an XML name repository to store these records nor to create "shareable" XML metadata in order to make local authority records interoperable within national and international cooperative XML authority databases. Finally, this method is a practical alternative for those libraries and institutions that do not plan to build an automated tool to extract names directly from the XML records, which so far has not proven to be a reliable alternative.

The University of Tennessee Libraries' Name Authority Challenge

At the beginning of 2007, the University of Tennessee Libraries (UTL) transferred the creation of descriptive metadata for digitized manuscripts from the Digital Library Center (DLC) to the Technical Services Department. After archives were scanned and digitized in the DLC, digital surrogates of the manuscripts were created using the Text Encoding Initiative (TEI) schema. TEI is a markup language for representing structural and conceptual features of texts. It is used primarily for the encoding of documents in the humanities and social sciences and, in particular, in the representation of primary source materials for research and analysis. Files in TEI were sent to the cataloging department to be transformed into rich descriptive metadata using the Metadata Object Description Schema (MODS), UTL's selected schema for digitized manuscripts.[1]

As a requirement of using MODS, catalogers have to use controlled vocabularies to assign access points to the records. Soon after receiving their first batch of TEI encoded records, catalogers encountered serious difficulties in assigning personal names to the access points of MODS records. The following were the main problems:

- **Difficulty in finding names in TEI records in the LCAF.** This problem occurred because either the record did not exist in the LCAF or because names in the TEI records could not be matched with names in the LCAF because of the lack of sufficient biographical information in the TEI records to identify individuals. For example, proving that the individual in the TEI record was the same one listed in the LCAF was difficult because no data other than name were given.
- **Inconsistency in the establishment of names not found in the LCAF.** When names were not found in the LCAF, different catalogers assigned different headings for the same person, depending on the form of the name given in the manuscript. Entering the same individual's name in many different ways can create a serious problem for future discovery and access.
- **Difficulty in differentiating individuals with similar names within the same collection.** Many people whose names appeared in the manuscripts in the UTL collection shared the same or similar names with relatives mentioned in the collections. Distinguishing between two or more individuals with similar names became difficult because little, if any, biographical data were provided in the manuscripts. To make matters worse, individuals sometimes were called only by nicknames or had very commonly used names, which were difficult to differentiate from other similar headings. These factors created confusion for catalogers and made the process of differentiation almost impossible.
- **Uncertainty about how to handle misspelled names and other typographical errors in the TEI transcriptions.** Sometimes errors were made in transcribing names from the digitized image to the TEI files. Catalogers did not know whether to go back and fix the misspelling by editing the TEI record or to create an access point using the form found in the manuscript, even if it were a misspelled form. Different decisions made by various

catalogers brought more inconsistency to the access points.

The problems in assigning personal headings to the access points of MODS records demanded an effective method to handle name authority control in the UTL's digitized collections.

Literature Review

Libraries and other institutions seeking to support name authority control in XML metadata have tried various approaches. Some have proven more successful than others, but none has consistently been implemented for XML documents. Some commonly mentioned approaches in the library literature include Metadata Authority Description Schema (MADS), MARC Extensible Markup Language (MARCXML), Encoded Archival Context (EAC), OCLC Linked Authority File (OCLC LAF), and the Automated Name Authority Control (ANAC). Most of these authority initiatives for non–MARC metadata are designed to handle authority control at the local level; only a few try to do so at the national level. Some are XML schemas for authority elements created for use in conjunction with particular XML bibliographic schemas. Others are conversion tools that convert MARC into XML records. Some authority initiatives claim to be automated, but usually these are really semiautomated approaches that apply a mixture of manual and automated approaches to generate authority control.

XML Schemas for Authority Elements in Non–MARC Metadata

In the early 2000s, the LC created MARCXML, an XML schema that can be used for authority purposes and is based on and very similar to MARC 21. It was first presented by the Information Technology Section at the IFLA conference in Glasgow in 2002. In a recent report of that meeting, McCallum states that "a key characteristic of MARCXML is that it produces an exact equivalent of the MARC 21 record so that roundtrip conversion to and from it is lossless. This schema has been widely used and is the basis for the international standard for an XML version of the MARC structure that Danish standards have proposed."[2] In summary, she concludes that MARC/XML "provides a basis for evolution while maintaining standardization."[3]

Later in 2005, the LC developed MADS, another schema for authority elements, but this one was created to be used in conjunction with MODS, a particular bibliographic schema. As in the case of MARCXML, MADS also has a strong relationship with the MARC 21 authority format. Guenther describes one advantage of MADS: "Because MADS is derived from the MARC 21 Authority format, which has been used for more than 30 years, its underlying model is well-established [and] a MODS description could link to a MADS description to eliminate redundant information."[4] She also mentions disadvantages: "Since MADS has not yet been widely implemented, it could still be considered experimental, and wider experience using it may result in refinements to the schema."[5]

EAC (www.library.yale.edu/eac) is another schema for authority elements created to be used in conjunction with a bibliographic schema, Encoded Archival Description (EAD). EAC started as an original effort from a group of archivists who met in Toronto in March 2001 to create a model for name authority control in archival materials. The initiative, still in the beta phase, is currently managed by an international group of archivists and Yale University. Thurman explains that EAC allows "archivists to encode information [in XML] about the creators and context of creation of archival materials, and to make that information available to users as an independent resource separate from individual finding aids."[6] He notes that EAC "development is not yet complete, and it has so far been implemented only experimentally."[7] In the effort to create an XML encoding standard for archival authority control, Pitti concludes that "there are many difficult intellectual, technical, cultural, linguistic, and political challenges to be addressed in order for the effort to be successful. While all of the challenges are significant, the political challenges stand out as particularly difficult."[8]

The MARC Extensible Markup Language Document Type Definition (MARCXML DTD), which is not the same as the MARCXML schema, is an older schema format for XML created by the LC. It started in the mid-1990s as an SGML DTD that supported the conversion of data from MARC Authority to SGML (and back) without loss of data. In the early 2000s, as technology developed and changed, the MARC SGML DTD became converted to MARCXML DTD. McCallum states in her report that this method "yielded very large DTDs since [XML DTD] is naturally verbose, and the tagging approach mandated a DTD element specification for every MARC subfield or coded character position."[9] An entry in Wikipedia summarizes the problems with DTD, noting that it is limited because "it has no support for newer features of XML, most importantly namespaces; uses a custom non–XML syntax, inherited from SGML, to describe the schema; and lacks expressiveness [because] certain formal aspects of an XML document cannot be captured in a DTD."[10] Nonetheless, even through its limitations, MARCXML DTD is still used and is kept available in the MARC 21 website. The reason some keep using it is that "several users have stated that they find it appropriate for certain applications, especially those needing extensive validation of records."[11]

Libraries that decide to use any of

these schemas to encode their XML authority data first will have to decide which names they want to extract from the XML records to be used as access points. This process has to be done manually because XML authority schemas do not extract information directly from XML records, but only encode it. The chosen names are then turned into access points, for which research is required. Next, the information is encoded into the desired XML authority schema. After this is done, a local XML name repository will need to be built and sustained to store and retrieve these authority records in XML. The problem with relying on XML name repositories for authority control is that catalogers frequently do not have the technological background to build or sustain the repositories. For these tasks catalogers often will have to rely on the library's programmers, who have competing responsibilities such as technical support or database and catalog maintenance. Hiring a programmer to work exclusively with the technical services department may be an option, but can be very expensive. For these reasons, the use of schemas for metadata authority control may not be the best solution for some libraries.

Conversion Tools from MARC to XML Authority Schemas

Another option for metadata authority control involves taking names that appear in the LCAF (in MARC format) and converting them into XML schemas using conversion tools between MARC and XML. Some of these tools involve automation, while others do not. An example of an automated conversion tool between MARC and XML is the MARC Tool Kit. This tool

> provides converters for transforming data from MARC 21 to MARC-XML and back, including character set conversion to and from Unicode. These converters can be downloaded from the MARC website and used by others in their own systems where they can also shape them to their own data and needs. [This] conversion software was developed by Bas Peters in the Netherlands and made available by him as open source software. It is in part adapted from an extensive set of programs for manipulating MARC 21 data.[12]

The LC sees these transformations provided from the MARC 21 maintenance agency as "being valuable to the community to help maintain the savings and interoperability built up through use of a common format."[13]

Maps and crosswalks between MARC and XML are other types of conversion tools used to translate authority data from one schema to another. These tools use manual approaches and, for this reason, require more effort on the part of the cataloger. The number of this type of conversion tool parallels the number of XML schemas. Some include conversions from MARC to Dublin Core and Dublin Core to MARC, others from MARC to MODS and vice versa. Almost all XML schemas have a crosswalk to convert their schemas into MARC metadata or from MARC to XML. An assessment published in *Online Libraries and Microcomputers* reveals some of the common challenges faced when using crosswalks and maps.[14] This analysis reports that

> there is often not a one for one mapping between fields in different metadata schemes. This means that many fields may need to be mapped into fewer fields (or vice versa). There can be a loss of granularity in metadata descriptions that may result in poorer searching. Many specific metadata schemes are targeted to a specific subject or type of material. When converting to another scheme there may be a loss of specificity and granularity. In metadata mapping one may want to parse through free text data to extract relevant data to extract for a more detailed scheme. This is difficult, time consuming and fraught with error because of variations in actual content. . . . How does one handle subfields and indicators (e.g., MARC) when mapping to systems that do not support the same detail? How should subfields from more complex metadata schemes be delimited in less complex metadata schemes? How does one map and handle local control numbers? Without the transferring of local control numbers there may be later problems in a shared database for updates, deletes and overlapping records.[15]

The idea of converting MARC authority records into records that use the local XML schema sounds appealing, but this method creates double work for the library. Converting authority records from MARC to another metadata schema requires translation of records plus the construction of an XML name repository to support the records. Many of the manuscript names do not exist in the LCAF, so locally established headings will have to be created for these names following the construction format of headings in the LCAF. Following the same construction format keeps consistency between locally created headings and those exported from the LCAF so that the headings look the same and index the same way.

If many headings have to be locally established in XML schema following the rigorous LCAF standards,

then libraries may find establishing the headings directly in the LCAF more worthwhile because other libraries can benefit from this authority work. This approach can also save the time necessary to convert names to another schema and to build a database to manage them. For these reasons, relying on conversion of authority records from MARC to XML may not always the best approach to support name authority control in XML metadata.

Endeavors for Cooperative Searchable XML Name Databases

Since the early 2000s, libraries and other institutions have attempted to create a national searchable XML name repository. Shared XML name repositories try to harvest name authority data from different sources distributed throughout the country and make it interoperable between different institutions. One example of such an attempt is the OCLC Linked Authority File Project (http://alcme.oclc.org/laf), an endeavor between the Open Archives Initiative and the OCLC. The Linked Authority File (LAF) was developed in 2002, hosts a shared server containing LC authority records and potentially authority records supplied by others, and is intended to provide Web-based accessed to interactive and automated authority records. This national name repository periodically uploads names from the LCAF and presents them in both MARC and MARCXML formats.

Even though the LAF's original intention was to harvest names from different sources besides the LCAF, this has not being done yet. When asked if the LAF plans to harvest authority data from other sources besides LCAF, an OCLC Research representative replied that "no further development of the system itself is planned."[16] No explanation was given on why the LAF only harvests authority data from the LCAF and not from other sources, but this may be due to the difficulty of making authority metadata interoperable between different institutions, a common problem faced by cooperative, inter-institutional databases. Given to the lack of promotion, this initiative is fairly unknown and, consequently, has not been widely implemented.

The Linking and Exploring Authority Files project (http://xml.coverpages.org/leaf.html) was an attempt to create a cooperative searchable XML database for authority names for the European community. It was created with the purpose of being accessed by anyone, regardless of affiliation, who might be interested in name authority files from European manuscripts. This three-year project (2001–4) was cofunded by the Information Society Technologies Program of the Fifth Framework of the European Commission.

Linking and Exploring Authority Files (LEAF) sought to develop a system model that uploaded name authorities—distributed through local servers of participating European organizations—to the central LEAF system. Authorities then were converted and stored into EAC schema, with authorities that belonged to the same entity being automatically linked. To have a network where those linked records could be applied, LEAF was integrated into a search engine called Manuscripts and Letters via Integrated Networks in Europe (MALVINE). MALVINE "is a search engine that harvests databases which provide information about letters written by famous persons that are kept in different European institutions."[17]

After being integrated into the MALVINE search engine, the linking process of LEAF proved not to be reliable. Kaiser and colleagues stated that

> it is inevitable that in some instances the linking process will produce incorrect results. Records describing two different persons might be automatically linked because they do not contain enough discriminating information. On the other hand, two records representing the same person might not be linked because they do not share an identical name form. Recollecting the main purpose of library authority records—the disambiguation of persons described—it may be argued that those records leading to wrong links are not sufficiently rich in content to serve their original purpose.[18]

The project ran as a funded test for thirty-six months, ending in May 2004. Thereafter it was left as an integrated part of the MALVINE search engine, where it is still used because it is seen as "highly relevant [content] to the cultural heritage of Europe."[19]

In theory, using a shareable XML name database sounds like a great plan for libraries and institutions that already have built a local XML name repository because records can be uploaded by one entity and shared between different institutions. In reality, experience has shown that this approach does not work because for metadata to be successfully harvested by a national cooperative repository, all locally created authority metadata needs to be "shareable." Shareable metadata are metadata that need to follow a set of standards to be interoperable between different institutions. The standards needed to create shareable metadata have not yet been established because of the lack of cooperation between different institutions. Pitti states, "As economically and professionally desirable as cooperative, shared authority control, and biographical, historical description is, successful realization will require standards and systems that are collaboratively developed, administered, and maintained. These

standards and systems will have to serve both individual and shared interests. Successfully balancing competing interests will require a great deal of patience, goodwill, and intelligence."[20]

Automated Endeavors to Support Name Authority Control in XML

Other projects have sought to solve the problem of addressing authority control in XML records on a local scale by implementing automated processes to extract and detect possible name access points in XML records. For example, in 2003, the Digital Knowledge Center (DKC) at John Hopkins University explored the application of automating metadata generation for name authority control.[21] To achieve this purpose, the DKC created a tool called the Automated Name Authority Control (ANAC). This automated metadata generator applies an established algorithm to identify LC–authorized names for each name in descriptive metadata records. Patton and colleagues stated that "The main reason for undertaking ANAC was to develop a tool that would reduce the costs associated with introducing name authority control to metadata [because] relying exclusively on human catalogers would be substantially more expensive and time consuming."[22] After evaluating the tool, the authors determined that the automated system was not sufficiently reliable in many cases. They added, "Even though ANAC could be a valuable complement [to authority control], it was never anticipated that it would entirely replace the human effort."[23] The authors concluded that the most effective and cost-efficient workflow would couple ANAC with human oversight.

Another attempt for automated generation of name authority control in digitized collections was suggested by French, Powell, and Schulman.[24] They introduced the concept of approximate word matching similar to the approximate string matching techniques traditionally used in detecting variant names in databases. This approach detects variable forms of strings in names through clustering algorithms and then groups the strings together under a standard form. The authors observed that, even though this automated clustering approach can reduce human effort by half, a certain amount of human effort will always be required to verify the output, thus this approach is semiautomatic.

Although systems created for the "automated" generation of name authority control claim to be automated, they are not completely so. They are really semi-automated approaches because they will always rely on human intervention for the process to work properly. Because of the need for human intervention and the high cost of creating such an endeavor, systems for the so-called automated generation of name authority control may not always be the best approach to support XML name authority control in many libraries.

Designing a Practical Approach at UTL

After reviewing the library literature and analyzing the advantages and disadvantages of the different approaches to support name authority control in XML metadata, UTL decided that none of these approaches was appropriate for the local situation. Because of time and funding constraints and lack of technological support, UTL decided to design a different approach that would be customized for UTL. Several points needed consideration. First, the new authority control method had to be completely sustainable by the UTL catalogers. Sustainable in this context meant the method had to be cost effective and use a level of technology with which the catalogers were comfortable. Because the cataloging department could not hire a local programmer, they ruled out building a local XML name repository and decided to capitalize on existing staff knowledge instead. Second, the authority control had to be achieved within a reasonable amount of time. Because this work is time consuming, priorities for which names to establish and which ones to leave out had to be set from the beginning. These priorities will be referred to from now on as "establishment criteria." The reason for using establishment criteria is that searching, verifying, and establishing each name found in the TEI records would be impossible because they number in the thousands. The criteria would determine the cases in which names would be searched and established in the LCAF. Third, the new authority method had to support a level of quality if UTL wanted to keep materials by the same creators together. Given these considerations, UTL decided to use a manual approach that keeps taking advantage of the largest name authority file available—the LCAF.

The method UTL implemented integrated all the previous points. The process is described in detail in the following section. Briefly, authority control is performed as soon as the DLC sends the TEI transcriptions to the cataloging department. Authority control, then, is performed before records are cataloged, and only by one person to avoid future inconsistencies in establishing names. The person chosen to perform this task is one of the catalogers, who had previous experience creating authority records through the Name Authority Cooperative Project (NACO). After authority work is finished, headings are stored in a Microsoft Excel shareable spreadsheet. As soon as the spreadsheet is ready, catalogers are notified and TEI records are sent to them. The catalogers then have the necessary resources to catalog and create rich, descriptive MODS records with the least amount of effort.

The detailed process followed at UTL consists of the following steps: The librarian charged with authority control receives a batch of one hundred to three hundred TEI transcriptions with their digital images from the DLC. The files usually originate from many different collections in the university archives. The authority control librarian devotes approximately two weeks (full time) to authority work for this batch of records. After receiving the TEI files, the authority librarian opens and browses the files using XML Pad. First, she groups the files by collection name. She identifies the collection to which a TEI file belongs by looking at the tag <collection> under the <sourceDesc> in the TEI file or by looking at the second portion of the TEI identification number. In the example 0012_000060_000337_0000, "60" identifies the collection to which the TEI belongs. This is illustrated in figure 1. (The zeros serve as fill characters and are omitted by the authority librarian when noting these data in working files.) By grouping TEI files by collection, catalogers will start working with all TEI files in one collection before moving to the next one. The logic behind this approach is to become familiar with the finding aid of a single collection by reading it once instead of having to read it many times. Another reason is that names in TEI files within one collection usually relate to each other, making keeping track of family and other types of relationships easier.

Within each TEI file, the authority librarian browses the following sections to search for names:

- Title section, to retrieve the names of senders and recipients.
- Body section, to find names that "pop out" as important access points. If summary sections are provided, the librarian should browse through them as well.
- Signature section, to determine the preferred form of heading for the sender.

Names found in these sections become crucial because they will form access points for the MODS record, the equivalent of the fields 1xx (main entry fields), 6xx (subject access), and 7xx (added and linking entry fields) in MARC records. The authority librarian should pay attention to variant forms of headings as well.

The authority librarian makes a list of the names found, as well as of their variant forms, along with the record number of the TEI where the names were found. By recording this number, the librarian can later retrieve the exact location of these names in case more information is needed. In going through the rest of the TEI files, the authority librarian may encounter the same names, as well as other new names or variant forms of names, and keeps a list. This process is illustrated in figure 2.

After browsing the TEI files in one collection, the authority librarian looks at the names gathered so far and

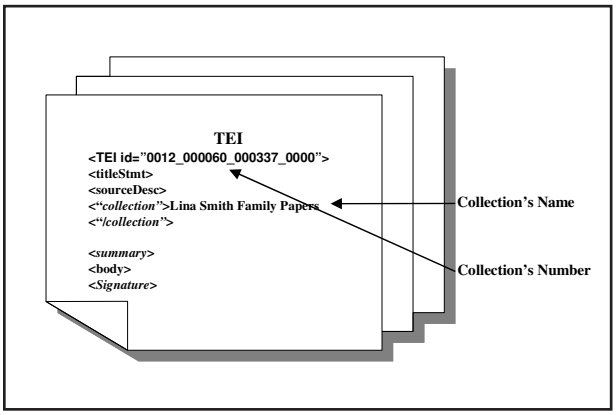

Figure 1. Sections to Browse in the TEI File to Determine the Collection to which the TEI File Belongs

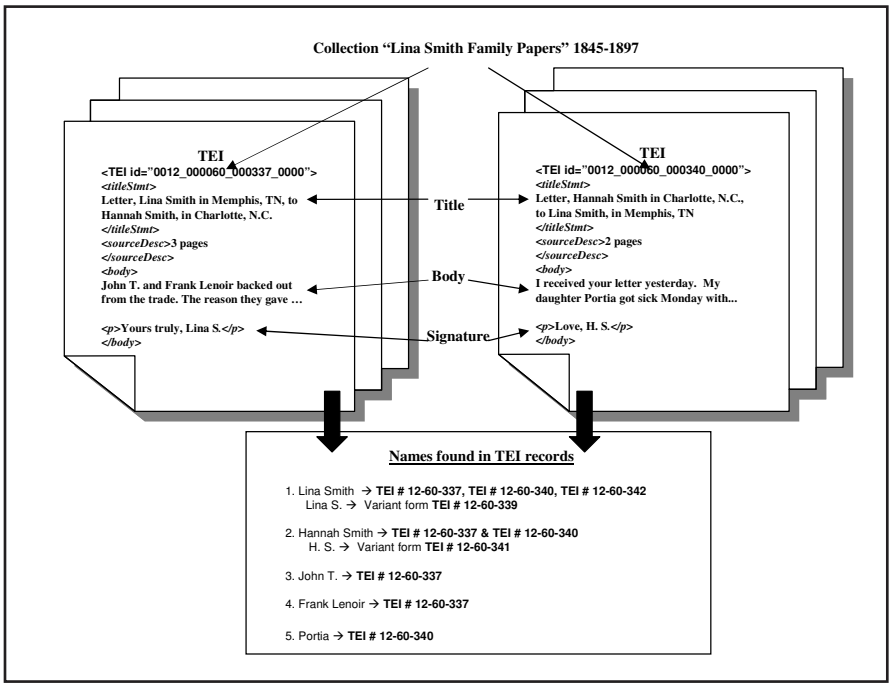

Figure 2. Sections to Browse in the TEI File to Get Names and How to Properly Keep a Record of Them in a List

compares them to see if some names have been mentioned more than once using the same form or a variant one. She also counts the number of times each name is mentioned in different TEI files.

The establishment criteria are then applied to names in that particular collection. These criteria help determine which headings will be searched, verified, or established, and which ones will not. UTL developed establishment criteria that worked well in most situations. Names mentioned in at least three separate TEI files are searched in the LCAF and established if not found. The same process applies to names mentioned in the "Title" section of the TEI files as senders or recipients and to names that have a collection with their name (this can be checked in the "Collection" section of the TEI file). The one exception to the establishment criteria is the handling of names for prominent historical individuals. Because they are likely to appear in the LCAF, they are also searched. Figure 3 illustrates application of the establishment criteria.

For names that will be established according to the criteria, the authority librarian returns to the TEI files in which they were found. This is done by using the TEI record number that was noted on the list. When retrieving the TEI records, the librarian browses the text around the area where the names were found to get as much information—stated directly or indirectly—about the person as possible. Examples of useful areas to browse in TEI files are the "Title" section, which gives the date and place a letter was sent, and the "Body" section, which may provide information on people's roles, relationships, and so on. The authority librarian annotates this information, along with the variant forms of the name found. The result might look like this:

> Jacob Breck, Jab Breck, Jacob B.; sender of letter from Franklin, TN in 1864 to Philadelphia, Penn; judge; wife Lizzie.

These brief factual data will provide a general idea of who this individual was and when he or she lived. The data found can be expanded later through further research in outside sources.

After all biographical facts available in TEI files have been annotated, the authority librarian then consults various research tools such as finding aids. The University Archives, which own the original texts for the TEI files, have created finding aids, many of which are online. Tennessee state and county archives also may contain related finding aids. These tools may provide information on the person's time period, family, place of residence, and more details. This information will be used by the authority librarian to place this person in context, see with whom he or she associated, and differentiate the individual from others with similar names when searching the LCAF.

If finding aids do not provide enough information about a person or are not available, the librarian searches other outside sources such as Google Book Search. This tool provides the ability to search sections within long texts of reliable resources that are freely accessible online. Other useful and freely accessible websites for historical biographical research include the Political Graveyard—A Database of Historic Cemeteries (http://politicalgraveyard.com), the state finding aids via the state library or state historical society websites, Genealogybuff.com, the Biographical Directory of the United States Congress (http://bioguide.congress.gov/biosearch/biosearch.asp), and the Civil War Rosters website (www.geocities.com/Area51/Lair/3680/cw/cw.html). The latter can be searched by soldier, regiment, and more. In addition, the authority librarian searches the Tennessee Genealogy and History Web (TnGenWeb.

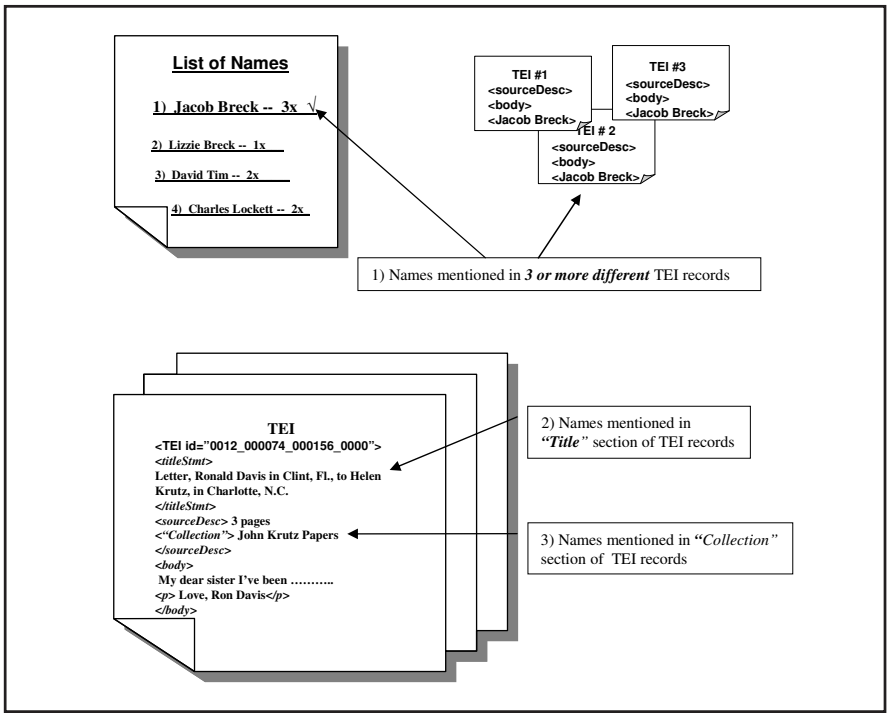

Figure 3. Application of Establishment Criteria for Names

org)—other states may have similar sources. Other fee-for-service genealogy databases are also available. When searching historical personal names online, a useful tip for best results is to search using very specific factual data. For example, if the only information available from a TEI transcription about "Lord Cornwallis" indicates that he was alive during 1791 and wrote from Blount County, these facts should be integrated into the search.

As relationships between individuals start becoming clearer, the authority librarian should illustrate the relationships using visual aids in addition to noting the information. Visual aids such as genealogical trees, arrows, and diagrams can prove useful to represent relationships between individuals. Visual aids are important for the authority librarian because she may need to consult these aids to create names, and they are important for the rest of the metadata team, who will assign the names as access points in the MODS records. Understanding relationships among the individuals in the digitized transcriptions is crucial to create useful access points for the records.

After gathering enough data about a particular individual, the librarian searches the name in the LCAF. At this point, she will have enough biographical information to distinguish that individual from others with similar names in the LCAF. If the heading is found in the LCAF, the authority librarian copies and pastes the heading into a local Excel spreadsheet, along with its cross-references and notes. The spreadsheet serves two purposes: It builds a local database of the established names found in UTL's digitized archives and provides catalogers with a narrower list of established names that appear in the TEI files they will catalog, saving them the time and effort of searching the LCAF. Sometimes, the authority librarian has located extra information not already mentioned in the LCAF about the individual listed. This extra information, such as biographical details or other variant forms of names, can be added optionally to the LCAF record in order to enhance it. This additional information can help differentiate this person from others with similar names in the future.

If the heading is not found in the LCAF, the librarian searches the OCLC Connexion Bibliographic File for records that used this name in any of their access points. To search in these areas of the bibliographic files, the authority librarian performs keyword searches using the index labels "au" (author) and "su" (subject). Searching in the OCLC Bibliographic File is a step required before establishing any heading in the LCAF. This search often leads to records that mention variant forms of this person's name as well as extra facts not discovered previously.

After searching the OCLC Bibliographic File, the authority librarian establishes headings that were not found in the LCAF using the biographical information gathered to this point. Headings can be established locally or nationally, depending on the institution's involvement with NACO. Libraries that are NACO members or part of a NACO Funnel Project have the option of making name contributions nationally. A NACO funnel project is a group of libraries who together are authorized to contribute name authority records to the LCAF. On the other hand, libraries that are not NACO members will not have the option of making national name contributions and will have to establish them locally. UTL has the option of making name contributions to the LCAF because it is a member of the Tennessee NACO Funnel Project. When establishing a heading, the authority librarian includes all the cross-references and factual data found previously in the research that may prove useful for the future. After establishing a heading in the LCAF, the authority librarian copies and pastes the heading into the local Excel spreadsheet with all the other LCAF names already found in OCLC.

After the chosen headings from one particular collection have been searched and established, the authority librarian browses the TEI files of the next collection, repeating the steps described above until all collections in the batch of TEI files are completed.

After names that met the establishment criteria have been searched or established, the lists of names that did not meet the establishment criteria remain. These lists are kept by the authority librarian in case any of the names need to be established in the future. Each list contains the TEI record numbers indicating where those names were found and can help retrieve the records if they are needed later.

After the authority librarian completes these steps, the authority work is considered completed. The tools and resources needed for cataloging metadata are then placed in a shareable department server. These include the digitized files in JPEG, transcription files in TEI, visual tools, and the Excel spreadsheet with the authorized name headings. The catalogers are then prepared to start creating MODS descriptive metadata with the least amount of inconvenience.

Implementing this authority control process before the rest of the metadata production starts solved the problems UTL initially faced when trying to assign name access points to MODS records without authority control. This approach to authority control solved both the difficulty in finding TEI names in the LCAF and the inconsistency in establishing names if they were not found there. Now that the authority librarian provides all the necessary authority work, the catalogers will not have to worry about searching these names in the LCAF or establishing them. The catalogers will

find the established forms plus their variants in a local, shared Excel list.

Placing authority control before the rest of the metadata process permitted the catalogers to focus on the rest of the description. It solved the difficulty of differentiating individuals with very similar names within a collection by providing useful biographical information. The use of qualifiers and other attributes in authority control also helped in this purpose. The provision of visual aids such as genealogical tables helped catalogers throughout the process of visualizing family relationships and helped to diminish confusion about similar names.

The problem of misspelled names and other typographical errors that occurred when transcribing names from the original text to the TEI files was also solved with this authority method. By receiving the TEI files with their digitized images as a first step, the authority librarian had the opportunity to catch transcription mistakes and fix them before the catalogers had the chance to discover them.

Assessing the Effectiveness of UTL's Approach

To assess the effectiveness of this approach, UTL decided to compare the metadata workflow before having authority control with the workflow after implementing authority control. UTL performed an informal assessment through a questionnaire, asking the six catalogers who experienced the first workflow without authority control to compare particular production aspects within both workflows. The questionnaire was distributed three months after the implementation of authority control into the metadata workflow and consisted of ten closed questions and one open question to provide suggestions. The questionnaire is presented in the appendix to this paper.

In the questionnaire, the six catalogers were asked if the speed of producing MODS records improved after the implementation of precataloging authority control. All six agreed that the speed of producing MODS records was higher after the implementation of authority control. When asked to estimate the number of MODS records produced per week before the implementation and the number produced per week after the implementation, they reported a much higher number of MODS records produced per week after the implementation. The six catalogers responded that before the implementation, an average of five or less records were produced per week; after the implementation, five catalogers reported an average of ten or more records produced per week and one cataloger reported six to nine records per week.

Catalogers were asked if the provision of authority control, before they began metadata work, freed them to concentrate on other important descriptive metadata tasks such as assigning subject headings, writing summaries, and analyzing the TEI record. To this question five of the six catalogers responded yes, the provision of authority control freed them to perform other important metadata tasks; one cataloger answered that it made no difference. Concerning quality of MODS records produced, all six agreed that the quality of MODS records improved after the implementation of authority control. Reasons for the quality improvement of MODS records were that more controlled access points were available than before the process changed, and that they were more consistent. Five of the six catalogers agreed that MODS records were more difficult to create before having the new-approach authority control. Reasons given to explain this difficulty before having the new approach were that there were inconsistencies in names established, distinguishing different persons with similar names was more difficult, and no visual tools were available to clarify relationships between individuals. Of the six, only one cataloger reported that the difficulty of creating MODS records was the same before and after the implementation of authority control.

Future Plans

While UTL's informal assessment demonstrated the effectiveness of this authority method in improving the MODS metadata production workflow, it also showed aspects that need improvement and issues that will need to be addressed in the future. In the suggestions at the end of the assessment, two catalogers showed concern about what will happen to the metadata workflow if the authority librarian leaves. To solve this, UTL will eventually need to expand and delegate authority control tasks to other members in the cataloging department so that authority control does not depend on one person's contributions. Initially, some authority control responsibilities, such as research tasks, can be delegated to members within the cataloging department. Eventually this responsibility can expand, with the catalogers creating personal authority records. They will need training either from the local authority cataloger who has NACO experience or through the closest NACO Funnel Project. Both alternatives would require initial time investment by the staff and institution, but this option could help make the workflow run more smoothly and to cover for the person performing authority work in case he or she leaves.

Another issue identified through the questionnaire was the increasing difficulty of searching names with many cross-references in Excel. As the number of names with cross-references increases, so does the difficulty in handling them effectively by the Excel software. Excel was not designed to

handle information arranged in thesauri format but primarily to handle numerical data. For this reason, commercial software that is better suited to handle cross-references will be needed in order to substitute for Excel. Thesauri software, which is software designed to build and edit thesauri headings, can manage cross-references very well and is cheaper than hiring a programmer to build an XML repository. Thesauri software is available in standalone packages and as database modules, which are integral parts of larger systems and need to run with them. Examples of popular standalone packages are MultiTes, Data Harmony, a.k.a. Classification Software, STRIDE, and Term Tree 2000. Examples of database modules are STAR and TheMa Thesaurus Manager for Oracle. Using thesauri software is an economical and attractive option to store and manage local authority names and one that UTL will begin to explore.

Conclusion

As evidenced throughout this paper, many libraries and institutions are looking for ways to turn necessary tasks over to machines, but experience suggests this is not yet possible for name authority control in XML metadata. The efforts created so far to achieve this goal, besides being costly and work intensive, have proved to be ineffective and unreliable. Most do not address the issue of how to extract or harvest names directly from the XML records and transform them into useful access points, but focus on how to encode the access points into XML authority schema. The few endeavors that have tried to harvest names directly from XML records have proved not to be completely reliable in their processes of matching and linking names, making them dependent on human effort.

In addition to not addressing how to select and extract access points from the XML records, most of these endeavors require labor-intensive encoding of authority data into XML schemas and, subsequently, the creation of a local XML name repository to store and manage these records. Building an XML name repository is a task that requires a high level of technological background most catalogers lack. For this reason a programmer will have to be hired to build a name repository, and this is an expensive approach not many libraries can pursue. Furthermore, creating authority data to be stored in a local repository will only benefit the local institution, causing inconsistencies and duplication of efforts between different institutions that try to set up access points for the same individuals. Initiatives that tried to avoid the duplication of efforts in name authority control—by creating a national XML name repository to share authority data and make it interoperable between different institutions—have not been successful because the XML authority data needs to be shareable to be interoperable between the national repository and the other institutions. To date, this has not been successfully achieved.

In contrast to these approaches, UTL's method to support name authority control in XML metadata is effective, reliable, and cost effective. It addresses the issue of extracting names directly from the XML documents and turning them into useful access points that can be shared nationally through the LCAF, thus avoiding duplication of efforts and benefiting all libraries who may share the same access points.

UTL's approach is simple and can be used by other libraries and institutions that face similar issues when trying to support name authority control in their XML metadata. Common problems such as inconsistency in the establishment of names, difficulty in differentiating individuals, and deciding which names to turn into access points can be solved by implementing this method before creating any descriptive metadata for digitized transcriptions. Regardless of the local XML schema used, this approach can be applied in the same way to different collections.

References

1. Library of Congress, Metadata Object Description Schema (MODS), MODS Schemas, www.loc.gov/standards/mods (accessed June 28, 2008).
2. Sally H. McCallum, "MARC/XML Sampler," *International Cataloguing & Bibliographic Control* 35, no. 1 (Jan./Mar. 2006): 4.
3. Ibid., 6.
4. Rebecca Guenther, "MADS," *Computers in Libraries* 27, no. 4 (Apr. 2007): 14.
5. Ibid.
6. Alexander C. Thurman, "Metadata Standards for Archival Control: An Introduction to EAD and EAC," *Cataloging & Classification Quarterly* 40, no. 3/4 (2005): 184.
7. Ibid., 199.
8. Daniel V. Pitti, "Creator Description: Encoded Archival Context," *Cataloging & Classification Quarterly* 38, no. 3/4 (2004): 217–18.
9. McCallum, "MARC/XML Sampler," 4.
10. Valid Documents: XML Semantics, DTD, en.wikipedia.org/wiki/XML (accessed July 5, 2008).
11. McCallum, "MARC/XML Sampler," 4.
12. Ibid., 5.
13. Ibid.
14. "Challenges and Issues with Metadata Crosswalks," *Online Libraries & Microcomputers* 20, no. 4 (Apr. 2002): 1–4.
15. Ibid.
16. Jeff Young, "RE: OCLC LAF question," e-mail to author, July 9, 2008.
17. Jutta Weber, "LEAF: Linking and Exploring Authority Files," *Cataloging & Classification Quarterly* 38, no. 3/4 (2004): 230.
18. Max Kaiser et al., "New Ways of Sharing and Using Authority Information: The LEAF Project," *D-Lib Magazine* 9, no. 11 (Nov. 2003), www.dlib.org/

dlib/november03/lieder/11lieder.html (accessed July 6, 2008).
19. LEAF Project Synopsis, www.crxnet.com/leaf/info.html (accessed July 7, 2008).
20. Pitti, "Creator Description: Encoded Archival Context," 218.
21. Mark Patton et al., "Toward a Metadata Generation Framework: A Case Study at Johns Hopkins University," *D-Lib Magazine* 10, no. 11 (Nov. 2004), www.dlib.org/dlib/november04/choudhury/11choudhury.html (accessed March. 27, 2008).
22. Ibid.
23. Ibid.
24. James C. French, Allison L. Powell, and Eric Schulman, "Using Clustering Strategies for Creating Authority Files," *Journal of the American Society for Information Science* 51, no. 8 (June 2000), www.cs.virginia.edu/papers/Using_Clustering.pdf (accessed March 27, 2008).

Appendix. Comparison of Metadata Workflow Before and After Implementation of Authority Control

1) Do you think the speed of producing MODS records was higher?
   a) Before the provision of authority control
   b) After the provision of authority control
   c) It was the same before and after
2) Do you think the quality of MODS records produced was better?
   a) Before the provision of authority control
   b) After the provision of authority control
   c) It was the same before and after
3) If you answered "after" to the previous question, why do you think the quality of MODS records was better after the implementation of authority control? Choose all that apply:
   a) Because there were more controlled access points
   b) Because access points were consistent between records
   c) Because records were more accessible to users
   d) None of the above
4) Do you think the production of MODS records was more difficult?
   a) Before the provision of authority control
   b) After the provision of authority control
   c) It was the same before and after
5) If you answered "before" to the previous question, why do you think it was more difficult to produce MODS records before the provision of authority control? Choose all that apply:
   a) Because there were inconsistencies in names established
   b) Because it was harder to distinguish different persons with similar names
   c) Because there were no visual tools available to understand relationships between persons
   d) None of the above
6) Do you think the provision of authority control for names in metadata records frees you to concentrate in other important tasks such as assigning subject headings, writing an abstract, or analyzing the TEI records?
   a) Yes
   b) No
   c) It makes no difference
7) Do you think the provision of authority control for names before MODS are produced improves the metadata workflow in general?
   a) Yes
   b) No
   c) It makes no difference
8) On average how many MODS records did you create per week before the implementation of authority control into the metadata workflow?
   a) More than 10
   b) 6–9
   c) 5 or less
9) On average how many MODS records did you create per week after the implementation of authority control into the metadata workflow?
   a) More than 10
   b) 6–9
   c) 5 or less
10) In which aspects of authority control would you like to see more improvement? Choose all that apply:
   a) Searching names in Excel spreadsheet
   b) Illustration of visual aids
   c) Time for authority control to be ready
   d) Other, please explain: _____
   e) None
11) Do you have any additional comments or insights regarding authority work for the metadata workflow? (For instance, recommendations for workflow, tools improvement, adjustments, and so on?)

Thank you for taking the time to answer the questionnaire!