Title: Western Name Authority File: A pilot regional name authority project

Authors: Anna Neatrour and Jeremy Myntti

Abstract

The prospect of authority control in digital libraries creates unique challenges. Digital library systems and software often do not support integrated authority control, which can create issues in consistency for personal and corporate names representation in descriptive metadata. Standard practice for library metadata is to use existing controlled vocabularies such as the Library of Congress Name Authority File, but what can be done if the personal names and corporate bodies in local or regional digital collections are not represented in the Library of Congress? As digital collection managers look towards providing metadata for regional and statewide shared repository systems and national digital collection aggregators like the Digital Public Library of America (DPLA), issues in digital collection authority control are magnified. This article explores the process in creating a shared regional authority file of personal names and corporate bodies existing in digital collection metadata records in several institutions throughout the Western United States. Steps in the process included reviewing data models, metadata collection, metadata deduplication and wrangling, vocabulary reconciliation, and data enhancement. Details on the process in making the Western Name Authority File accessible to the public and assessing project outcomes are included.

Introduction

The University of Utah J. Willard Marriott Library has a long tradition in supporting regional partners through collaborative digitization. With over thirty partners, such as the University of Utah Eccles Health Sciences Library, Park City Historical Society, the Utah Department of Heritage and Arts, and the Uintah County Library, the digital library program has grown greatly

over the years. Digital libraries in the Mountain West have a long collaborative tradition due to the presence of the Mountain West Digital Library (MWDL), which was founded in 2001, and was one of the 6 initial service hubs that launched with the Digital Public Library of America (DPLA) in 2013. In addition to encouraging regional collaboration on the development of digitization best practices, the service model of MWDL where larger institutions like the University of Utah provide digital library repository hosting for partners has provided great opportunities for growth, along with the development of lingering issues of metadata quality control as our supported digital collections grow larger.

Seeing our metadata aggregated for many years in both a regional and national context increased our awareness of inconsistencies of personal and corporate names, especially when facets are applied to aggregated metadata. When embarking on a new program of metadata assessment and remediation, examining and improving personal and corporate names was a first step in an attempt to make our digital library metadata more consistent and be better positioned to take advantage of the promise of linked data, especially the prospect of utilizing additional information about people and corporate bodies that could be leveraged for enhancement such as the Library of Congress Name Authorities or Wikidata.

The Western Name Authority File (WNAF) project developed after a long process of reviewing and testing methodology to clean up personal names and corporate bodies in our digital library. First, the Marriott Library developed a pilot project with Backstage Library Works to test reconciliation for names in XML based metadata exported from CONTENTdm, using a system previously developed for MARC based reconciliation. This process provided name and subject headings that had been matched against the Library of Congress Name Authority File (LCNAF) along with many different reports to review for possible additional matches and corrections (Myntti & Cothran, 2013). In a review of these reports, it was discovered that the majority of

names used within our digital library did not match up with an existing record in the LCNAF, creating the need to investigate local authority records for those individuals. A variety of additional methods of vocabulary reconciliation were also tested using a set of names that Backstage was not able to identify through their reconciliation process, identifying a few more names that were represented in external vocabularies, and further refining a list that could be used for further local vocabulary projects (Myntti & Neatrour, 2015). As these projects were developing, we were also reflecting on the fact that our challenges were likely to be shared by other members of the MWDL community. Many other institutions in the region were using CONTENTdm as their digital library repository system, which provided little in the way of built-in tools to support vocabulary maintenance and authority control.

After testing reconciliation methods with a subset of our digital collections we wanted to extend authority control options for existing partners hosted on our repository, as well as colleagues in the MWDL community. In addition to gaining additional expertise from colleagues for our project, we wanted to explore if the process for authority control for digital libraries could be made more efficient through regional collaboration. The next logical step was to begin the basis of a regional authority file that could be used as a mechanism to ensure better quality control into the future for our digital collections.

Literature Review

Authority control for bibliographic information has long-standing cataloging conventions as well as established collaborative processes in the form of the Name Authority Cooperative Program (NACO), as established by the Library of Congress Program for Cooperative Cataloging (PCC). Integrated Library Systems (ILS; e.g., Ex Libris' Alma, SirsiDynix's Symphony) and third-party vendors (e.g. Backstage Library Works, Marcive) have established methodologies for MARC-based authority control in an ILS, however the options for authority control in the digital

repository landscape are much more limited. Lee summarizes this situation by stating, “As the need for identity management has grown in recent years due to increased interest in preserving more types of outputs and the proliferation of online works, the limitations of the traditional process of name authority control are more pronounced. The creation of authorized name headings can be too slow and unresponsive as it relies on a body of work and a preferred form of the name” (Graham, Lee, Radio, & Tarver, 2018, p. 26).

Additional systems for managing name authority control have also been developed in recent years, including the International Standard Name Identifier (ISNI), developed by the ISNI International Agency to “assign to the public name(s) of a researcher, inventor, writer, artist, performer, publisher, etc. a persistent unique identifying number in order to resolve the problem of name ambiguity in search and discovery; and diffuse each assigned ISNI across all repertoires in the global supply chain so that every published work can be unambiguously attributed to its creator wherever that work is described.” Direct contributors to ISNI include national libraries, agencies, and ISNI also harvests NACO data via the Virtual International Authority File (VIAF) (ISNI International Agency, n.d.).

While national efforts for authority control provide key infrastructure for managing names at a large scale, there is still a need for local name authority management. The Shareable Local Authorities Forums and white paper, funded by the Institute of Museum and Library Services (IMLS) and hosted by Cornell University, provide an overview for issues of local authority control. The report detailed a variety of areas for sharable local authorities including minimum viable specification, data provider obligations, workflows, and the idea of reconciliation as a service (Casalini et al., 2018).

In response to the scalability issues and time-consuming aspects of traditional authority control, a new process, called “NACO Lite” has been proposed. In the Charge for PCC Task Group on Identity Management in NACO, the strategic direction for the PCC is articulated as “Provide leadership for the shift in authority control from an approach primarily based on creating text strings to one focused on managing identities and entities” (Program for Cooperative Cataloging, 2018). A major goal of the NACO Lite process includes lowering the barriers to complete this type of authority record creation by creating new minimum requirements for NACO authority records. Making NACO contributions easier is a laudable goal for the future, but while waiting for more official documentation and workflows from the PCC, institutions are developing their own methodologies to share the burden of authority control.

Learning how to provide NACO records and integrate that process into existing workflows can prove challenging, as seen in a case study provided by the University of Nevada, Reno, which embarked on NACO training and certification with catalogers and metadata librarians. Their solution blended teams of library staff through “a workflow that captures and funnels vital information to NACO-certified catalogers who can then use that information to create name authority records” (Miller & Hunsaker, 2018, p. 147). This solution makes the creation of NACO records easier through non-cataloger contributions to the LCNAF, but it still relies on specialized training and staff with the time to create and submit these records. (Miller & Hunsaker, 2018)

Lampert summarizes the current landscape for digital library repository managers engaging in authority control by stating, “some systems enable metadata creators to import locally created vocabularies or link to vocabulary services to access lists of terms” (Lampert, 2017, p. 166). Lampert continues by discussing the need for systems that make it possible to review and include new local terms which generally doesn’t exist in the workflows of current systems. (Lampert, 2017)

While large cooperative organizations are currently investigating how to make identity management work more accessible, institutions still face the issue of ensuring authority control and identity management for their own digital collections. Cultural heritage digital work also requires authority control, with further complications due to the fact that the names represented in cultural heritage materials are often not accompanied by additional metadata or information for name disambiguation. Digital collection repository managers have a variety of approaches to authority control, ranging from developing collaborative NACO workflows, keeping localized spreadsheets of authority information, and developing databases of entities associated with digital collections.

The University of Nevada, Las Vegas (UNLV), developed a unique Linked Open Data Navigator for their Southern Nevada Jewish Heritage Project, which allows users to interact with linked data triples, or the subject, predicate, object and object information for names and relationships in a visual prototype: http://lod.library.unlv.edu/nav/jhp/. The navigator is built with a framework of CONTENTdm for metadata management, TemaTres for managing controlled vocabulary, OpenRefine for LOD transformation, and OpenLink Virtuoso (Lampert, 2017).

Veve notes problems with developing name authority files for XML (eXtensible Markup Language) documents, and shares a process that involves extracting data from national sources when possible, and saving local names in an excel spreadsheet designed to be used as an internal source of name authority for Text Encoding Initiative (TEI) encoded manuscripts at the University of Tennessee (Veve, 2009).

Challenges in providing name authority control for local digital collections are explicated by Dragon, using a case study of postcards where there were issues in developing subject

headings for entities such as buildings. The challenges were, “(1) the complexity of work arising from the form and subject matter of the materials digitized, (2) the volume of work created by a high ratio of new authorized headings per bibliographic description, and (3) the inefficiency perpetuated by the lack of actual authority data in the repository database” (Dragon, 2009, p. 185).

The University of Denver has examined issues of archival authority control, sharing plans for the development of a shared authority tool. They conceptualize this tool by situating it in user communities in the Rocky Mountain region which may include small cultural organizations with an interest in people, family relationships, and historical institutions. The primary goal of the tool would be to “provide a highly focused window into our own locally established records about the creators and subjects of our collections, especially those that may not exist in any other linked open data set.” Many archives are in a similar place with their authority work as the Special Collections Department at the University of Denver, with locally developed authority information that does not yet exist in another linked open data set (Crowe & Clair, 2015).

The need for local name authority is not limited to cultural heritage materials that are held by special collections, as detailed in a case study about the University of North Texas Name App, which develops local authority records which can also be connected to external authority files. The need for local authority data in the University of North Texas Name App was articulated after a study of their institutional repository, which showed that a small percentage of faculty had authorized name forms. Texas A&M University has implemented VIVO to manage faculty names and research, saying that “the system aggregates heterogenous, authoritative data from internal and external databases, and allows the faculty to manage or control their own scholarly narratives by contributing authoritative data” (Graham et al, 2018, p. 26). Integration with ORCiD

is a methodology being explored for similar local authority needs at the University of Arizona (Graham et al., 2018).

Oregon Digital, a unique digital repository program that combines digital collections from the University of Oregon and Oregon State University, manages a shared linked data authority file called OpaqueNamespace. The process of developing OpaqueNamespace involved extensive work in metadata migration during a systems migration. While external vocabularies were used and Oregon Digital adds names to external vocabularies as part of their process, “it is inevitable that a local list will always be maintained” (Simic & Seymore, 2016, p. 312).

While efforts are beginning to investigate workflows to make contributions to large scale external authority files through programs like NACO, the current state of name authority control for digital repository managers leaves them with gaps both in repository system support of integrated authority control, and with the infrastructure to support customized local authority control lists. While many individual institutions develop their own local authority file, this type of program is likely out of reach for many smaller institutions who do not have the staff time or technical infrastructure in place to develop local name authority solutions.

Western Name Authority File Grant Funded Project

In 2016, the Marriott Library was awarded a planning grant from the Institute of Museum and Library Services (IMLS) under the National Digital Platform program. This planning grant included a four-stage project to investigate the creation of the Western Name Authority File (WNAF), a controlled vocabulary of personal names and corporate bodies used in digital collections metadata records from multiple institutions in the Western United States. This four-phase project included:

1. Investigation: In the first phase of the project, metadata from project partners was collected and evaluated as potential data for inclusion in the WNAF. Multiple data models were explored to help best represent this type of authority data given the data available for the project. Baseline statistics were gathered to assess the changes in discoverability at the end of the project.
2. Testing and evaluation: Phase two included exploring many different open source tools that could be used to create, manage, maintain, and provide access to the data in the WNAF. A large-scale metadata wrangling project was conducted to bring together all of the different types of data submitted by partner institutions.
3. Pilot implementation: Once a tool was selected for the project, a pilot implementation was conducted to fully evaluate the data and software. Workflows for creating and maintaining data were developed. Additional workflows for creating NACO records were developed to share data from the WNAF with the LCNAF.
4. Assessment: Project outcomes were measured in order to discover the impact of this type of regional controlled vocabulary. A toolkit has been developed for other institutions to replicate the project and implement a similar project using their own data.

Data Model Review

During our early meetings with project partners (Brigham Young University; Oregon Digital; University of Denver; University of Nevada, Reno; Utah State Archives; Utah State University), we discussed the fields that would be most useful to have in our regional vocabulary system. We agreed early on that we wanted not just authorized forms of names and variants, but also additional information such as institutional holdings for names as well as information about digital collections where names are represented. We also discussed possible data models and issues in authority work by exploring the representation of authorities in BIBFRAME v.1 vs the Agent/Role in BIBFRAME v.2 ("Bibliographic Framework Initiative," 2018), SKOS (W3C, 2009),

OWL (OWL Working Group, 2012), LCNAF (Library of Congress, 2018), and EAC-CPF (Technical Subcommittee on Encoded Archival Standards (TS-EAS), 2018).

The team ultimately decided upon EAC-CPF for a variety of reasons. The names in WNAF are largely drawn from digitized special collections, making a standard developed by the archives community extremely well-suited to the project. Also, we examined the Social Networks and Archival Context (SNAC) project, which uses the EAC-CPF standard to accomplish linking between archival collections at a large scale (Larson, Pitti, & Turner, 2014), which provided a useful model for similar work with multiple institutions.

Metadata Investigation

At the beginning of the project, we requested that our partner institutions send metadata from their digital collections containing historical local or regional names that could be included in the WNAF. Since there were a variety of systems being used by our partners, we let them choose how they sent this data in order to simplify the process for them. We received data in a variety of formats, including plain text files with lists of names, Java Script Object Notation for Linked Data (JSON-LD), Comma Separated Values (CSV) or Tab Separated Values (TSV) files containing the full metadata from a collection, and spreadsheets containing names along with a wide variety of extraneous local data. We had to work separately with each different type of file in order to standardize and compile the data into one large dataset. This work was completed using tools such as Notepad++ for working with simple text files, Microsoft Excel and LibreOffice Calc for working with spreadsheets and CSV/TSV files, and an online tool for converting JSON-LD to a tab delimited text format (Data Design Group, n.d.). While converting each of the different types of file into the standardized format, we created a common set of core fields to retain for each name:

- Name as used in the digital collection
- Alternate form of the name (if available)
- Institution submitting the name
- Collection containing the name (if available)
- Metadata field containing the name (if available)
- Type of name (personal name, corporate body, family name -- if available)

After compiling all of the names into one master spreadsheet in Microsoft Excel, we had a dataset with over 500,000 lines of data. There were many duplicate names in this dataset, so the next step was to deduplicate the data based on exact matches. When deduplicating, we wanted to make sure to retain all of the information connected with the names, so we combined the institution, collection, and metadata field into one standardized field (institution;collection;field). When two names were deduplicated, we were then able to append these multiple fields together, separated by [space][dash][dash][space] (e.g., USHS;Classified Photos;Person -- UU;UAIDA;Creator). Using a standardized format like this made it possible to separate out this data later on in the project.

The first step of de-duplication took place in Microsoft Excel using an "if/then" formula (e.g., =if(A1=A2,"DUPLICATE","")). This formula compares the contents of cell A1 and A2. If they are exact matches, then the formula will print "DUPLICATE" in the new cell. If not an exact match, the new cell will be left blank. Once potential duplicates were identified, they were reviewed and combined into one row where appropriate, making sure to retain all data describing the institution(s) and collection(s) where the names exist.

After this initial deduplication, we were able to condense the dataset from over 500,000 names to approximately 76,360 names. This dataset was much more manageable to work with for the

pilot project. We created a Google Sheet containing this data so that multiple people could work on it and we could share our progress with our project partners.

Since our partners are spread across multiple states (Colorado, Nevada, Oregon, and Utah), we added a new field for the state name based on the institution(s) that submitted that form of the name. We were able to start doing some analysis of the data to find which names have been used most often within particular states, multiple institutions, and multiple collections. We found some examples of collection and institution overlap, which we expect to grow over time, as we continue to deduplicate and further research the names in the dataset. Of the names that were initially gathered and deduplicated, we found:

- 7357 names were used in more than one collection/field (9.6%)
    - 13 were used in more than 20 collections/fields
    - 80 were used in more than 10 collections/fields
    - 6795 were used in 2-5 collections/fields
- 1484 names were used in more than one institution (1.9%)
    - 1360 in two institutions
    - 110 in three institutions
    - 11 in four institutions
    - 3 in five institutions
- 271 names were used in more than one state (0.35%)
    - 267 in two states
    - 4 in three states

The largest number of names types were 62,381 personal names, with 10,706 corporate bodies, and 3,273 unknown. In the dataset, 1091 names were single words, over 2400 were

cross references, and over 500 were written in the format of first last, instead of last, first, as is traditionally standard formatting.

Total names submitted from partner institutions

- Brigham Young University - 30,535
- Utah State Historical Society - 12,138
- University of Denver - 16,608
- University of Utah – 7533
- Oregon Digital - 4170
- Utah State Archives – 3657
- Utah State University - 2067
- University of Nevada, Reno - 1277

After the data had been cleaned up with most duplicates resolved, we created a workflow for our student research assistant to reconcile the data against the LCNAF. There are many established workflows and reconciliation services for this type of task, so we repurposed the work of Matt Carruthers and Jennifer Wright from the University of Michigan (Carruthers & Wright, 2015), which was chosen based on our previous experiences testing reconciliation methods (Neatrour & Myntti, 2015). Carruthers and Wright provide a detailed method for name reconciliation by using VIAF, scoped for Library of Congress name authorities. This process avoids the occasional downtime or access issues that occur sometimes with id.loc.gov, and it was easy to train students on the process of applying an extracted operation history to a spreadsheet in OpenRefine (Carruthers & Wright, 2015). By using this workflow, the student assistant was able to process many spreadsheets of names through the reconciliation service to identify potential matches with the LCNAF. After the reconciliation was complete, the student

would review the matches to identify those that were most likely accurate and those that weren't. With 55,314 personal names reconciled against LCNAF, we found that 7382 of the matches were valid (13.35%), 9251 of the matches were not valid (16.72%), and 38,681 didn't have any potential match in the LCNAF (69.93%).

Based on the reconciled data, we were able to identify many names in our digital library as well as those of our partners that were not using the current authorized access point according to the LCNAF. We generated reports for all of our partners similar to Table 1 listing the name that the institution is using, the form of name in the LCNAF, and the digital collection where the name has been used. We encouraged our partners to review these possible changes and make updates in their local repositories as necessary. We also made these updates in the University of Utah's digital library. Based on the data that we reviewed in this phase of the project, we were able to update 14,133 metadata records in University of Utah and Utah Department of Heritage and Arts collections. Two major examples of these changes included updating over 40 variations on "Savage, C. R. (Charles Roscoe), 1832-1909" and over 400 variations of "Shipler Commercial Photographers" (see Table 2 for examples). Providing users with one form of these names to search within our repository has helped to improve discovery by collocating all items related to specific names together.

Table 1: Reconciled data to clean-up in local repository

Table 2: Example of Shipler name variants

Software Testing and Evaluation

After evaluating and combining the metadata for the regional authority file, the team explored possible software solutions for storing the data in a web accessible format. At an early stage of the process, the project partners and the PIs agreed on a common set of core fields that the

vocabulary should contain and discussed a variety of potential schema for WNAF. Being able to eventually build a database that would accomplish similar linking between personal and corporate names along with associated digital collections and digital items is a long-term goal for our project.

Choosing EAC-CPF as our data model caused some additional complications for our pilot project when we wanted to investigate open source software to store our data. With limited funding for custom development at this stage of the project, we were limited to software that has been developed for more general projects, and we didn't have the time or personnel to create a custom solution for our vocabulary. In addition, since a regional authority file needed to be referenced by our partner institutions, we required a system that had a web-based discovery layer, with an infrastructure that was not tied to a particular system such as CONTENTdm or Samvera.

We developed a rubric for software testing that looked at the following components for each system:

- Project name, documentation, and web site
- Technical support considerations in our local environment (installation and ongoing support)
- Software type (backend, middleware, framework, complete solution)
- Linked Open Data publishing capabilities
- Batch import and export support
- Search functionality (browse and advanced search)
- Data model(s) supported
- Testing notes

We coordinated with the library's Digital Infrastructure Development department to have versions of vocabulary management software installed on a sandbox server for testing. As we evaluated the software, a number of issues surfaced for our project. Many vocabulary management systems such as TemaTres (Ferreyra, n.d.) assume a thesaurus-like list of terms for the vocabularies they support. Since WNAF had a more granular model for information to be potentially associated with each term, a traditional glossary or hierarchical thesaurus structure wasn't suitable for the project. VocBench (Stellato, Turbati, Fiorelli, & Lorenzetti, n.d.) was in between versions while we were evaluating software for our project, with a new release just issued in fall of 2018.

In the end, the most important functionality that we needed to evaluate open source software on was support for customized vocabularies. After testing several solutions and closely reading the documentation for solutions we were not able to test, we initially settled on CollectiveAccess ("CollectiveAccess," n.d.) for our vocabulary solution. CollectiveAccess is an open source collection management solution, but the functionality we were interested in primarily was the vocabulary management feature, which allowed the development of custom vocabularies and batch upload. The built-in structure for managing entities in CollectiveAccess matched up closely with the vocabulary metadata fields we had decided on with our project partners. However, we still had some practical implementation concerns for our pilot with CollectiveAccess, as we would have needed to gain additional expertise in the web-based administrative functions of the system, and even after several troubleshooting attempts, we had difficulty uploading our data through batch upload through the provided spreadsheet template. Testing the various types of vocabulary management software also left us with the impression that we would likely end up with a solution that would be good enough for a pilot, but full implementation for WNAF would involve custom software development in order to develop a

system that contains batch editing features and support for EAC-CPF as well as supporting collaborative workflows for ongoing vocabulary management.

Pilot Implementation

At the same time as we were examining software for WNAF, we were also investigating software for a new digital exhibits program for our library. We realized that while Omeka S (Corporation for Digital Scholarship, n.d.) wasn't on our initial list of software to test for the WNAF project, it had many of the features we were looking for, including support for custom vocabularies, an API to potentially support reconciliation, the ability to publish data as JSON-LD, a search and discovery layer, and editing functionality.

As a first step, we took the EAC-CPF Schema and ran it through the online conversion tool ReDeFer (Garcia, n.d.) to generate an RDF/XML file Omeka S would recognize for import. Once the vocabulary terms were in place, we proceeded to the next phase of testing.

We next tested OmekaS batch import capacity. The Omeka S CSV import plugin helped us to manage bulk imports by providing an easy way to view previous jobs and undo them when we noticed any quality control issues that required additional clean-up work in the CSV files we were uploading. While one initial attempt showed us that it was possible to upload a CSV file with over 50,000 vocabulary entries to our instance of Omeka S, we ultimately chose a more distributed approach of uploading our 60,567 vocabulary terms in batches of 7,000 to 12,000. This also allowed us to more easily pinpoint potential issues with unicode encoding errors which would cause the CSV import functionality in Omeka S to break. To prevent this, we loaded each CSV into OpenRefine for one last quality control check before upload, and used OpenRefine's customized facet by Unicode char-code to set aside any names with encoding errors, with the

plan to fix and upload them at a later date. The pilot dataset is available at https://exhibits.lib.utah.edu/s/wnaf/page/welcome, along with search tips and a form for suggesting new personal and corporate names.

For our pilot project, OmekaS gave us the functionality we needed to make our dataset searchable and available via an API and we are investigating methods of making it available through bulk download as JSON-LD. However, there are a few limitations of this approach that we would need to investigate further if the project were to move to a full implementation. We were initially hopeful that the provided REST API from Omeka S would give us a good solution for reconciliation with OpenRefine for our project partners. However, we ran into several errors when we investigated this functionality after our vocabulary was uploaded. In addition, in the future we would like to explore providing our vocabulary in a triple store to better enable us to visualize the relationships between the entities in collections and representation in institutional holdings.

Integrating WNAF with NACO

From the reconciliation work to find names that were already in the LCNAF, we were able to identify multiple projects using WNAF data that could potentially add new records to LCNAF through the NACO process or update existing LCNAF records. For names that successfully reconciled with a record in the LCNAF, we isolated authorized access points that did not have a death date. We were able to identify 186 names with a birth year and no death year where the person was born before 1918, so they were most likely no longer living. We were able to find death dates for 165 of these names to add to the LCNAF records. There were an additional 203 names without a death date, but the person would have been between 70-100 years old. We identified the death dates for 89 of these records. As of this writing, we have been able to update 67 of these LCNAF records and plan to update the additional 187 records in the near

future. There were 195 names that didn't have a death date and based upon the assumption that the person would likely still be living and under 70 years old, we did not proceed with additional research at the current time.

Another area for NACO work was to identify names that did not currently have a record in the LCNAF, but were good candidates for including in the national authority file since they had been used in more than one institution or else they had items in at least three collections. This included over 2500 potential records that could be created and added to the LCNAF. Since it takes a great deal of time to complete the research and record creation for a name to go through the full NACO process, we created a workflow that would allow a student research assistant to conduct some basic research and then pass that information to a NACO-trained metadata cataloger for final review and record creation (Myntti, 2018). As of this writing, the student research assistant has completed research for 531 names, 15 of which have been reviewed by the metadata cataloger and submitted to the LCNAF.

Since the NACO process is time-consuming, identifying these types of projects can help to create new records as they are needed. As evidenced by the workflow to create new NACO records, it can be easy to devote student time towards research but finding the time of a NACO-trained cataloger to verify and finalize the records can be difficult as they manage the myriad of other projects assigned to them.

Project Assessment

One aspect of metadata change that we wanted to measure was changed facets for personal names and corporate bodies in our regional digital collections aggregator, the Mountain West Digital Library (MWDL), as well as in the Digital Public Library of America (DPLA). To accomplish this, we received additional assistance from the Marriott Library Digital Infrastructure

Development department in the development of a stats script that would query the DPLA API for the presence of personal names and known variants and misspellings. The script is available on the Marriott Library's GitHub repository. Names to be queried were placed in a SQLLite database, which provided a quick and simple method of taking advantage of Structured Query Language database features, and the Python Requests library was used to query and return values using the DPLA API (Reed & Neatrour, 2017).

The assessment script was run two times, once at the beginning of the project and once at the end of the project after the WNAF file was developed and additional metadata corrections were implemented. There were 582 changes out of 4087 sample names during the lifetime of the pilot project. These changes represent a smaller number of partners since not all partners have had their metadata reharvested by MWDL and DPLA since these changes to the local repositories have been implemented.

When observing the changes in this sample set of names, there are several facets of aggregated metadata for personal names that show definite improvement. For example, there was previously a wide variety of name variants, including misspellings for the term Shipler Commercial Photographers, such as 117 records for "Shipler Comm. Photog." and 544 for "Shipler Commercial Photography". After the metadata was corrected and reharvested, the facet for the creator term was clustered around the term "Shipler Commercial Photographers", with 20,932 items for that term in 2018, up from 17,173 previously in 2017.

Partners who contributed their data to the pilot WNAF commented that it "was a valuable exercise for us to consider our own vocabularies and how they can function across geographically distant collections" (A. Hunsaker, personal communication, June 22, 2018). Other partners are drawing upon templates and workflows developed for NACO work related to

WNAF, in particular the University of Oregon, which is developing a method to “integrate the creation of personal name authorities with our institutional NACO contributions, our local controlled vocabulary manager (opqauenamespace.org), and WNAF” (Seymore, personal communication, September 7, 2018). Utah State University plans on using WNAF as another authority source in metadata workflows, alongside the LCNAF (L. Woolcott, personal communication, July 3, 2018).

Lessons Learned and the Future of WNAF

While many tasks associated with metadata cleanup, enhancement, reconciliation, and developing controlled vocabularies can be automated, it is important to carefully consider the variety and scope of manual work associated with building a regional controlled vocabulary. While we were planning for a certain amount of manual work associated with WNAF, particularly in the area of reviewing results from vocabulary reconciliation, we eventually realized that we had not anticipated many aspects of the manual work at the beginning of the project. Some of the manual tasks which took more time than expected include needing to standardize metadata produced in a variety of partner institutions and systems, the work of deduplication, researching near matches, formatting data for systems testing, and the demands of quality control.

Now that we have completed the pilot project for investigating and implementing the WNAF, we are looking towards the future of the project and how this can be sustainable going forward. While a lot of time and effort has been spent throughout this pilot project to make sure our dataset is fairly clean, there is still a substantial amount of manual review that would need to be completed to identify additional duplicate names. With over 60,000 names in the project at this point, it will take a large amount of time to do a thorough review of all of this data. There are also over 10,000 names that we decided to remove from the pilot project due to extra manual

review and time constraints to complete the project such as names where we only had a last name (e.g., Mr. Smith) or names that were only expressed as initials (e.g., A.C.B.).

Basic workflows have been developed for our current project partners to be able to submit new names to the WNAF. These workflows still involve some manual work on the backend in order to ingest the names into our current system. In order to bring more partner institutions on board with the project, we will need to develop better methods for reconciling existing data against names already in the WNAF and simplify the process to add new names or link new collections to existing names.

Workflows for updating names that exist in both the WNAF and LCNAF as well as for adding new names to the LCNAF have been developed. While these workflows are functioning, there is a backlog of manual review that has to be completed before all of the names that have had adequate research can be submitted through the NACO process. This is another example of the extensive manual time requirements that can be taken up with this type of authority work.

In order to investigate how we can improve the WNAF project and make it more useable by other digital library metadata creators, a follow-up research project is in the initial planning stages to discover how other institutions are tackling this issue. By gathering qualitative data about the status of authority control in digital repositories, we hope to better refine our method for authority control and make the WNAF a resource that has a wider impact on the usage of these types of regional controlled vocabularies.

We have recognized that while Omeka S has worked well for the pilot implementation of the WNAF, it does not provide all of the necessary features and functions for maintaining this type of vocabulary in the long term. In conjunction with another project at our library, we are planning

on investigating the use of a triple store for this data and how we can make better use of the information and relationships that have been exposed through the WNAF project.

In order to complete all of these future tasks, we will need to find the resources necessary to devote large amounts of time to making the full implementation of the WNAF possible. We have learned many lessons through this pilot project that will help us to be more informed in a full implementation once we have the resources to do so.

A full implementation of WNAF as a regional authority control project for digital collections would be able to provide digital collections metadata specialists with a centralized place to engage with authority work when creating metadata for new digital objects. By building upon shared regional knowledge, digital libraries would be able to realize greater efficiencies in determining who among our collection of names is likely to be significant enough to engage in more in-depth research, which would more effectively position WNAF as a feeder source for NACO work within the region. Characteristics such as birth dates and occupation are important for name disambiguation, in addition the presence of location information associated with these names, curated by WNAF, may also assist in developing richer information about the entities described in our digital collections. It is the authors' intention that a full implementation of WNAF could additionally serve as a model for other institutions who routinely provide metadata that is aggregated in regional or national contexts.

Engaging in a regional authority control project for digital libraries helped us realize the depth of the work that lies ahead in improving representation of entities in our digital collections. Being able to work with metadata provided by our partners also made us realize that many of the issues in our own metadata are common across other digital repositories. Over time, as we invest more research time to surface additional connections and engage in further

disambiguation and deduplication, we expect to build beneficial regional descriptive metadata workflows as well as reinvigorate our collaborative contributions to our national authority files as well.

Author Acknowledgements:

A collaborative project like the Western Name Authority File is not possible without support and contributions from many people. The authors wish to thank our partners on the project for contributing their metadata and expertise; Kayla Willey and Rebecca McKown from Brigham Young University, Liz Woolcott and Andrea Payant from Utah State University, Sarah Seymore and Julia Simic from the University of Oregon, Amy Hunsaker from University of Nevada, Reno, and Gina Strack from the Utah State Archives.

At the University of Utah Marriott Library, we wish to thank Jacob Reed and Curtis Mirci for support with software development and testing, Nicole Lewis for providing comments on a draft of this article, and Ambra Gagliardi for grant development and support.

We would also like to thank The Institute of Museum and Library Services for funding the project.

References

Bibliographic Framework Initiative. (2018). Retrieved from https://www.loc.gov/bibframe/

Carruthers, M., & Wright, J. (2015). *Breaking the bottleneck: Automating the reconciliation of named entities to the Library of Congress Name Authority File*. Presented at the American Library Association Midwinter Meeting, Chicago, IL. Retrieved from https://deepblue.lib.umich.edu/handle/2027.42/138107

Casalini, M., Chew, C. N., Cluff, C., Durocher, M., Folsom, S., Frank, P., & Gatenby, J. (2018). *National strategy for shareable local Name Authorities National Forum : White paper*. Cornell University. Retrieved from http://hdl.handle.net/1813/56343

CollectiveAccess. (2018). Retrieved from https://www.collectiveaccess.org/

Corporation for Digital Scholarship. (2018). Omeka S. Retrieved from https://omeka.org/s/

Crowe, K., & Clair, K. (2015). Developing a tool for publishing linked Local Authority Data. *Journal of Library Metadata*, *15*, 227–240. https://doi.org/10.1080/19386389.2015.1099993

Data Design Group. (2018). Convert JSON to CSV. Retrieved from http://www.convertcsv.com/json-to-csv.htm

Dragon, P. M. (2009). Name authority control in local digitization projects and the Eastern North Carolina postcard collection. *Library Resources & Technical Services*, *53*(3), 185–196. https://doi.org/10.5860/lrts.53n3.185

Ferreyra, D. (2018). TemaTres. Retrieved from http://www.vocabularyserver.com/

Garcia, R. (n.d.). ReDeFer. Retrieved from http://rhizomik.net/html/redefer/

Graham, S., Lee, D. J., Radio, E., & Tarver, H. (2018). Who is this: Moving from authority control to identity management. *AALL Spectrum* 22(5), *2*4-27. Retrieved from http://epubs.aallnet.org/i/963996-aall-spectrum-may-june-2018-volume-22-number-5/25

ISNI International Agency. (n.d.). How ISNI works. Retrieved from http://www.isni.org/how-isni-works

Lampert, C. (2017). Looking at linked open data from a digital asset management perspective. *Journal of Digital Media Management*, *6*(2), 161–173.

Larson, R. R., Pitti, D., & Turner, A. (2014). SNAC: The Social Networks and Archival Context project - Towards an archival authority cooperative. In *IEEE/ACM Joint Conference on Digital Libraries* (pp. 427–428). London, United Kingdom: IEEE. https://doi.org/10.1109/JCDL.2014.6970208

Library of Congress. (2018). Library of Congress Name Authority File (NAF). Retrieved from http://id.loc.gov/authorities/names.html

Miller, D. M., & Hunsaker, A. J. (2018). Extending Name Authority Work beyond the cataloging department: A case study at the University of Nevada, Reno Libraries. *Library Resources and Technical Services*, *62*(3). http://dx.doi.org/10.5860/lrts.62n3.136

Myntti, J. (2018). Western Name Authority File Project NACO Workflow. Retrieved from https://sites.google.com/site/westernnameauthorityfile/project-work-plan/naco-workflow

Myntti, J., & Cothran, N. (2013). Authority control in a digital repository: Preparing for Linked data. *Journal of Library Metadata, 13*(2-3). 95-113. https://doi.org/10.1080/19386389.2013.826061

Myntti, J., & Neatrour, A. (2015). Use existing data first: Reconcile metadata before creating new controlled vocabularies. *Journal of Library Metadata*, *15*(3–4), 191–207. https://doi.org/10.1080/19386389.2015.1099989

OWL Working Group. (2012, December 11). Web Ontology Language (OWL). Retrieved from https://www.w3.org/OWL/

Program for Cooperative Cataloging. (2018, May 22). Charge for PCC Task Group on Identity Management in NACO. Library of Congress. Retrieved from https://www.loc.gov/aba/pcc/taskgroup/PCC-TG-Identity-Management-in-NACO-rev2018-05-22.pdf

Reed, J., & Neatrour, A. (2017). Simple DPLA data gathering script. Retrieved from https://github.com/marriott-library/simple-dpla-data-gathering-script

Simic, J., & Seymore, S. (2016). From Silos to Opaquenamespace: Oregon Digital's migration to Linked Open Data in Hydra. *Art Documentation: Journal of the Art Libraries Society of North America*, *35*(2), 305–320. https://doi.org/10.1086/688730

Stellato, A., Turbati, A., Fiorelli, M., & Lorenzetti, T. (2018). VocBench. Retrieved from http://vocbench.uniroma2.it/

Technical Subcommittee on Encoded Archival Standards (TS-EAS). (2018). EAC-CPF Homepage. Retrieved from https://eac.staatsbibliothek-berlin.de/

Veve, M. (2009). Supporting Name Authority Control in XML Metadata: A practical approach at the University of Tennessee. *Library Resources & Technical Services*, *53*(1), 41–52. https://doi.org/10.5860/lrts.53n1.41

W3C. (2009). SKOS Simple Knowledge Organization System. Retrieved from https://www.w3.org/2004/02/skos/