Linking People: Developing Collaborative Regional Vocabularies -- Abstract

For libraries taking the first step into Linked Open Data (LOD), using controlled vocabularies is an essential part of creating new data structures linking people, places, collections, and digital objects together. The Western Name Authority File (WNAF) will be a first step in collaboratively analyzing existing vocabularies, developing a data model, exploring infrastructure, and testing workflows that could be used throughout the Mountain West Digital Library (MWDL) network of partners. Building on existing work at the University of Utah's J. Willard Marriott Library of reconciling digital collection metadata fields against existing controlled vocabularies, this project would explore creating a regional vocabulary in an open and shareable format using a process that can be replicated at other institutions.

Currently in the MWDL and at local institutions, name variants provide users with unnecessary additional search options. A central name authority file like the WNAF can provide an essential reference tool for catalogers and metadata librarians. In addition, many MWDL partners are currently using vendor-based software that is not linked data compatible. The WNAF will provide a LOD project for librarians in the region to develop, opening up new opportunities for training and workflow development. The WNAF will serve as a model for other Digital Public Library of America (DPLA) hubs wishing to investigate methods of local LOD authority control.

WNAF will be developed during this planning grant by the University of Utah, along with partners Utah State University and Brigham Young University, with assistance from project participants from the Utah State Library, Southern Utah University, the Utah State Archives, University of Nevada Reno, Boise State University, and the University of Oregon. The PI's for the project will also be guided by an advisory board, consisting of representatives from DPLA, Stanford University, and others yet to be determined, which will meet virtually during the four phases of the project.

This planning project will take place over two years, with four distinct phases of investigation. The first stage will collect and evaluate personal names metadata from the project partners and participants, and review data models for the WNAF. With a data model selected, the next phase will center on testing open source software functionality, ease of use, and collaborative support for developing a pilot WNAF dataset. The full pilot test will result in the creation of the pilot WNAF dataset, including data enrichment and collaborative workflow development and documentation. After the pilot dataset is created, workflow changes will be assessed, as well as the changes in metadata from the partners and participating institutions.

While the grant activities are going to be limited to a few MWDL partners, the project is designed with wider collaborative possibilities firmly in mind, and we will be releasing workflows, documentation, and the pilot WNAF dataset for reuse by other institutions as a toolkit. Upon conclusion of this project, we would have information in place to consider the requirements, workflow, and costs associated with wide scale implementation of a regional LOD controlled vocabulary for personal and corporate names in the Mountain West region.

Linking People: Developing Collaborative Regional Vocabularies

1. Statement of need

Across many institutions, controlled vocabularies for personal names and corporate bodies (hereafter names) are maintained in siloed information environments, for example as locally developed text fields within a particular CONTENTdm repository. Descriptive metadata work in the Mountain West region could benefit greatly from a shared controlled vocabulary system for names, as librarians and archivists could draw on shared expertise about people or corporations that are notable regionally, but not likely to be within the scope of national vocabularies like the Library of Congress Name Authority File (LCNAF). Expressing controlled name vocabularies in a shared infrastructure that is Linked Open Data (LOD) compliant will help metadata catalogers in the region become more familiar with LOD technologies, as well as provide an infrastructure to visualize new connections between the entities represented in digital library collections. One of the biggest potentials for the future of LOD is the decentralization of authority data, making it possible for the reuse of a regional vocabulary on a global scale.

The four principles of LOD as described by Tim Berners-Lee are: 1) Uniform Resource Identifiers (URIs) should be used as names for things, 2) the URIs need to be created in the Hypertext Transfer Protocol (HTTP) so that they can be accessed by others, 3) useful information in the Resource Description Framework (RDF) standard is provided at the URI, and 4) links to other related URIs are included to help the user discover other information.[1] In order for a URI to exist to represent a piece of data, there needs to be a controlled vocabulary containing information related to that "thing."[2]

Many large-scale library efforts in LOD have focused on bibliographic data, through BIBFRAME[3], BIBFLOW[4], Linked Data for Libraries (LD4L)[5], and OCLC's linked data efforts with Worldcat[6]. For regional and local repositories managing digital collections, the WNAF will provide an important first step in developing collaboration in a LOD environment.

Heath and Bizer have said that in "order to make it easier for applications to understand LOD, data providers should use terms from widely deployed vocabularies to represent data wherever possible."[7] A standard practice in libraries is to use the LCNAF as the main controlled vocabulary for names. However, in many local digital collections, there are many names that are not

---

[1] Heath, T., and Bizer, C. (2011) Linked data: Evolving the web into a global data space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. San Rafael, CA: Morgan & Claypool. Available at http://linkeddatabook.com/editions/1.0/

[2] Berners-Lee, T. (2006). Linked Data. Available at http://www.w3.org/DesignIssues/LinkedData.html

[3] http://www.loc.gov/bibframe/

[4] https://www.lib.ucdavis.edu/bibflow/

[5] https://wiki.duraspace.org/pages/viewpage.action?pageId=41354028

[6] http://www.oclc.org/research/themes/data-science/linkeddata.html

[7] Heath, T., and Bizer, C. (2011) Linked data: Evolving the web into a global data space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. San Rafael, CA: Morgan & Claypool. Available at http://linkeddatabook.com/editions/1.0/

represented in the national authority file since they do not necessarily fit within the scope of the LCNAF, and are only significant within a small location or region. Those local or regional names are more suited to be created and maintained within a controlled vocabulary specific to that geographic area.

Patricia Dragon discusses the complexities in creating authority records for local names when there are no occurrences in national bibliographic databases like OCLC's Worldcat. Some of the difficulties that arise with this are the different rules that the Library of Congress has for types of headings that may fall under either the name authority rules or subject authority rules. Another challenge is identifying the correct form of the name to use when there are multiple forms that are current or historical and there are no reference sources available that can assist with making a judgement about the name. Dragon mentioned that the volume of potential names that could be submitted to an authority file with digital projects can increase exponentially, making it difficult to contribute those names to the LCNAF. Dragon also speaks to the challenges of authority work for digital collections, which usually does not include an authority record in a controlled vocabulary, making it difficult to provide any context about the name or variant forms of the name.[8]

In talking about creating authority records for a local controlled vocabulary, Veve says that while it may be useful for a local XML based controlled vocabulary be created, it is often difficult for catalogers to use these types of vocabularies since they don't have the technical expertise to work with XML data and they end up relying on programmers to support the database. Veve reported that in working with manuscript collections, many of the local names do not appear in national authority files like the LCNAF. These names must then be constructed in a like manner as LC name authority records, but they are not assigned to a controlled vocabulary that can support ongoing access and maintenance. Veve also states that locally created authority records need to be shareable, requiring that interoperable standards be used. This is often difficult and not often implemented because it can be difficult to have multiple institutions collaborate on this type of shared vocabulary.[9]

In a survey about metadata decisions made for digital library content completed by Zeng, Lee, and Hayes in 2009, 66% of respondents said one of their major concerns for their metadata was regarding the fields that should use a controlled vocabulary. These respondents had concerns about the choices in existing controlled vocabularies and the need to establish local vocabularies for data that didn't currently reside in an existing vocabulary such as the LCNAF.[10]

Many of the institutions contributing to the Mountain West Digital Library (MWDL) are using local instances of OCLC's CONTENTdm. Within CONTENTdm, there is a feature for creating basic

---

[8] Dragon, P. (2009) Name Authority Control in Local Digitization Projects and the Eastern North Carolina Postcard Collection. Library Resources & Technical Services 53(3). doi:10.5860/lrts.53n3.185

[9] Veve, M. (2009) Supporting Name Authority Control in XML metadata: a practical approach at the University of Tennessee. Library Resources & Technical Services 53(1). doi:10.5860/lrts.53n3.185

[10] Zeng, M., Lee, J., and Hayes, A. (2009) Metadata Decisions in Digital Library projects – Summary of an international survey. Journal of Library Metadata, 9(3-4). doi:10.1080/19386380903405074

controlled vocabulary lists. One difficulty for authority control using this feature is that each individual digital collection in CONTENTdm can have its own controlled vocabulary. There is currently no easy way to share the creation and maintenance of these controlled vocabularies with multiple institutions using CONTENTdm, even though the values in the vocabularies amongst the institutions may be very similar. The CONTENTdm controlled vocabulary lists also do not provide an easy way to maintain cross references or to distinguish which strings of data are in other vocabularies such as the LCNAF. These controlled vocabulary lists also have a 128 character limit for each term, making it impossible to store large strings of data. Rather than having controlled vocabulary lists in silos such as these CONTENTdm controlled vocabularies, this type of data is better served in a LOD environment where "decentralization, collaboration, localization, richness, and structure" are embraced.[11]

An example of LOD vocabularies in practice can be seen in a local installation of VIVO at Texas A&M, where it was determined that a local ontology was needed in order to create IDs for names. This ontology framework was created using RDF (Resource Description Framework) standards and OWL (Web Ontology Language) classes. As this project grows, they expect the local ontology to grow as well to be able to represent new information.[12]

As mentioned in the IMLS Focus Report discussing the National Digital Platform, "linked data is not something that can simply be bolted onto existing tools, practices and data; standards and shared protocols are needed to make it work, as well as fundamental work to rethink processes and workflows. The value is only realized when other resources are made openly available as linked data."[13] This project will help in achieving this goal by developing workflows using existing open source tools that other institutions can use to incorporate LOD in digital library metadata work.

The development of the WNAF will engage librarians in the region with a practical and needed LOD project. By providing a mechanism to collaborate on authority control work, regional digital library collection managers will be able to express personal and corporate names with greater consistency, improving search and discovery in local collections, and for metadata aggregators like the MWDL and the Digital Public Library of America (DPLA). Documenting the process of building the WNAF will serve as a model for institutions considering similar projects, thus contributing to the National Digital Platform.

2. Impact

This project is designed to provide a foundation for collaborative regional authority work, along with refining workflows and building infrastructure that could be further expanded on in the future for wider implementation. Many digital library managers in the Mountain West region are currently using systems that are not conformant with LOD principles. The project presents a

---

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

[12] Ilik, V. (2015). Cataloger makeover: creating non-MARC name authorities. Cataloging & Classification Quarterly, 53(3-4), 382–398. doi:10.1080/01639374.2014.961626

[13] https://www.imls.gov/sites/default/files/publications/documents/2015imlsfocusndpreport.pdf

practical educational opportunity for digital library managers to engage with an aspect of LOD, further improving knowledge and professional development opportunities in the region.

In addition, developing a pilot centralized resource for authority control provides an opportunity to create greater efficiency for authority control and descriptive metadata work within the region. Expressing names with more consistency across the multiple institutions who are harvested into the MWDL will improve access to resources for researchers, who will no longer have to explore multiple name variants in order to collect all resources about a particular person. While the scope of this grant being limited to only a few partner institutions will not result in resolving all of these issues, it will improve discoverability for names among the project partners, and could result in developing greater consistency for future projects.

To demonstrate issues with researching known people in the MWDL, here is a small illustration of name variants:

C. R. Savage, photographer of the American West, with photographs held by many different institutions, is expressed with these name variants:

- Savage, C. R. (Charles Roscoe), 1832-1909
- Savage, C. R.(Charles Roscoe),1832-1909
- Savage, C.R. (Charles Roscoe)
- Savage, Charles R.
- C. R. Savage (Charles Roscoe Savage and George Ottinger), Pioneer Art Gallery, East Temple Street, Salt Lake City, Utah
- Savage, C. R. (Charles Roscoe)
- Charles R. Savage
- Savage, Charles Roscoe
- C. R. (Charles Roscoe) Savage, photographer
- Savage, C. R.
- Charles R.

Frank Asahel Beckwith, newspaper publisher, photographer, and amateur anthropologist is expressed with these name variants:

- Beckwith, Frank Asahel, 1876-1951
- Beckwith, Frank
- Beckwith, Frank A.
- Beckwith, A. Frank

Creating greater consistency with names that are expressed in digital collections metadata will improve searching and discoverability at the national, regional, and local levels. Consistency in the forms of names entered in metadata also helps to gather related items and provides greater precision in results retrieved from a specific search. Making the pilot dataset, workflows, scoring model, and evaluation of software to support the creation of the WNAF widely available will provide a model for other digital libraries and state-based or regional digital library DPLA aggregators to consider for their own authority control projects. Gretchen Gueguen from DPLA says in her letter of support that "the approach to developing a regional

controlled vocabulary and use of linked data [...] could have significant and meaningful impact at institutions around the country, particularly at DPLA hubs. Consistency in the usage of subjects, names, and places in the metadata that is aggregated by DPLA is one of the largest detriments to the quality of our data set. The example set by the Marriott Library and the MWDL community in this project could provide a beneficial model to enable Hubs to not only create more consistent vocabularies, but to create them in a way that takes advantage of the power of linked open data."

In a recent informal survey, representatives from ten MWDL repositories expressed interest in receiving regular updates about the project. Fifteen representatives from these partners expressed interest in engaging with the activities listed in the project design below.

3. Project Design

This project draws on existing work of automating controlled vocabulary reconciliation that has been completed at the University of Utah's J. Willard Marriott Library. Seventeen collections have had names matched with the LCNAF through a combination of vendor provided reconciliation and with scripts in OpenRefine[14]. These collections have been updated with more accurate values, giving us a sample dataset of standardized local and regional names and an established workflow. These collections, along with additional data from partners, can form the basis of a regional LOD vocabulary for names. This previous work has been shared at national conferences, and will be featured in a forthcoming issue of the Journal of Library Metadata.[15]

This planning grant will allow the University of Utah to investigate, test, and pilot a workflow for developing the Western Name Authority File (WNAF), a shared vocabulary of names which could later be expanded to include additional partner institutions of MWDL. This project will provide a first step in standardizing names across MWDL partners, including libraries, museums, archives, and other cultural heritage institutions. Improving discoverability through the use of shared vocabularies will allow users to be more precise with their research within local, regional, and national discovery systems. Expressing the vocabulary as LOD provides the structure needed to make authority information open and repurposable.

A related Library Linked Data Project is currently underway at the University of Nevada, Las Vegas (UNLV).[16] We plan to coordinate with UNLV to make sure the two projects complement each other. The project being completed at UNLV will be focussing on advancing the knowledge of LOD in the region through workshops and exploring how collaborative vocabularies can be leveraged using their existing technologies while the project at the University of Utah will be investigating tools and conducting a pilot implementation and assessment of a regional controlled vocabulary.

---

[14] http://openrefine.org/

[15] Myntti, J. and Neatrour, A. (forthcoming) Use existing data first: Reconcile metadata before creating new controlled vocabularies. Journal of Library Metadata, 15(3-4). doi:10.1080/19386389.2015.1099989

[16] https://www.library.unlv.edu/linked-data

The DPLA Metadata Application Profile (v.4), is architected to allow for harvesting LOD-ready Uniform Resource Identifiers (URIs) which will provide the ability to link directly to LOD triple stores such as the one that would be created in this project. DPLA is currently harvesting URIs for spatial metadata from GeoNames with potential future plans to expand this service to other fields affected by local controlled vocabularies such as the one created in this project. The DPLA metadata team has been contacted in preparing this proposal has expressed excitement for the possibilities this project could have for MWDL and other DPLA-contributing institutions. A member of the DPLA metadata team will provide guidance throughout the project by being a member of the project's advisory board.

The project plan has the following four six month phases of work:

1. **Investigation:** The first phase of the project will collect and evaluate data from fields with controlled vocabularies from multiple partner institutions, such as the University of Utah, Utah State University, Brigham Young University, and other MWDL partners. We will collaboratively explore data models such as Encoded Archival Context - Corporate Bodies, Persons, and Families (EAC-CPF) standard[17], Simple Knowledge Organization System (SKOS)[18], Web Ontology Language (OWL)[19], BIBFRAME authorities, and similar projects such as Social Networks and Archival Context (SNAC)[20] that create and represent relationships between entities and collections. We will capture a baseline of analytics data by assessing how names in the selected vocabularies are currently discoverable in MWDL, DPLA, and Google. We will also collect exported collections data from partners in order to be able to document change after a regional controlled vocabulary is fully tested. We will adopt a data model for the vocabulary.
2. **Testing and Evaluation:** The second phase of the project will focus on the collaborative evaluation of open source software that can be used to create, maintain, and make available the data contained in a compiled regional controlled vocabulary, with collaborative authority control. We will develop a scoring model and evaluation criteria for reviewing software and infrastructure that could be repurposed by other institutions with similar projects. Towards the end of this phase, we will move forward with a pilot and full evaluation of selected software. Software and open source tools that will be evaluated include (but are not limited to) Protege[21], TemaTres Controlled Vocabulary Server[22], Skosmos[23], TMP2 (Terminology Management Platform) from AthenaPlus[24], and an Apache Jena RDF Triple Store[25]. These and other software that are identified will be evaluated using a scoring model developed during this phase of the project. Example criteria used to evaluate the software includes:

---

[17] http://www2.archivists.org/groups/technical-subcommittee-on-eac-cpf/encoded-archival-context-corporate-bodies-persons-and-families-eac-cpf

[18] http://www.w3.org/2004/02/skos/

[19] http://www.w3.org/TR/owl2-overview/

[20] http://socialarchive.iath.virginia.edu/

[21] http://protege.stanford.edu/

[22] http://www.vocabularyserver.com/

[23] http://skosmos.org/

[24] http://www.athenaplus.eu/index.php?en/212/tmp2

[25] https://jena.apache.org/

- Ease of use for creating and maintaining controlled vocabularies
- Batch importing of existing controlled vocabulary lists
- Batch editing of controlled vocabulary terms
- Ability to publish the vocabulary as LOD
- Advanced search and browse capabilities
- Ability to create URIs to represent the data values
- Data models supported
- Server requirements to install and maintain the system
- Capabilities for collaborating with multiple institutions
- Availability of APIs to further develop the system
- Tools to visualize data

The University of Utah's Marriott Library recently purchased new virtual machine (VM) equipment and storage, which allows us to have access to storage and computing resources that will be used to install and test the different software options in this project. The hardware is located in a secure data center with additional off-site storage for tape backups. Ubuntu, an open source software platform, will be used as the server software. Since some software options such as Skosmos require a triple store to house the data for a controlled vocabulary, we will use an Apache Jena triple store that can be accessed by multiple other software options.

3. **Pilot Implementation:** With the assistance of a part-time student research assistant performing data entry and additional support tasks, we will harvest, standardize, reconcile, and import controlled vocabulary information into the software of choice and make this data available as LOD. We will enrich data with relationships and collections holding information, and incorporate additional research about the people represented in the vocabulary, for example providing birth and death dates where applicable. We will explore the possibility of setting up an OpenRefine reconciliation service for the vocabulary, and documenting the reconciliation process. We will document collaborative workflows and assess impact on work for the University of Utah and partner institutions, showing how existing workflows would need to change to incorporate regular use of the WNAF.
4. **Assessment:** The impact of the project will be measured by reviewing the information gathered about the existing metadata and current search results before the WNAF is created, along with statistics showing changes in user access to names after metadata has been updated. Data to be gathered and assessed includes, but is not limited to:
    - Exported digital collections metadata for those collections identified as having the greatest number of regional or local names, before and after the WNAF is created and workflows for updating metadata values are developed.
    - Document cross institutional variance of names (e.g. Smith, Joseph, 1805-1844 vs Smith, Joseph, Jr., 1805-1844)
    - Collect data on the number of alternate forms of names that can be added as cross references to a "master" record in either LCNAF or WNAF.
    - Sample faceting comparisons in MWDL and/or DPLA for names before and after the project and document change. We will explore using Primo's X-services queries to pull this data from MWDL, or DPLA's API to capture this data.

- Search results from known searches before and after the WNAF is created
- Percentage of names in LCNAF
- Percentage of names used in a single institution or used in multiple institutions
- List of names that may need to be disambiguated
- Number of relationships between names that can be identified

We will further assess the outcomes of the project by reviewing workflows, identifying training opportunities, and exploring the impact of the centralized vocabulary on users of local digital asset management systems, regional digital library collaboratives (e.g., MWDL), and national level digital library efforts (e.g., DPLA). We will develop a plan for expanding the controlled vocabulary to more institutions.

In the process of identifying names to add to the WNAF, there will be instances where names should be added to the LCNAF. Names that are used in multiple collections across multiple institutions would be the top priority to be submitted to LCNAF. Depending on resources and name usage, priorities for submission to LCNAF will be created for each institution.

4. Project Resources: Personnel, Time, Budget

As co-principal investigators for the project, **Jeremy Myntti and Anna Neatrour** will coordinate work on the grant, including data model assessment, software testing, workflow development, communication with project partners and the larger community, and evaluation and assessment. Myntti has expertise in authority control, digital project management, and developing efficient methods for creating and updating metadata. Neatrour has a background in collaborative digitization, metadata quality assurance, and grant project management, serving as project manager on both LSTA and IMLS grants in the past. Myntti and Neatrour have recently completed work at the University of Utah on a local authority control project that has provided a pilot set of data for this grant.

A temporary part-time **Student Research Assistant** (19 hours per week for 48 weeks) will be hired during the second year of the project to provide assistance with data entry and assessment tasks associated with the full pilot of the selected solution. The student will perform research needed to document linkages between names and library collections, assist with vocabulary development and maintenance, test workflows, and assist with assessment activities towards the end of the grant.

This project will require the assistance of **Curtis Mirci**, an Application Development Programmer at the University of Utah's Marriott Library. Mr. Mirci will complete the majority of his work on this project in the early part of the testing and evaluation phase (second phase) where he will install and configure different open source software tools on the library's sandbox server for testing. He will also help in the pilot implementation phase (third phase) to make sure that the software selected in the evaluation phase is available for all those who need access both internal to the University of Utah as well as all contributing institutions. Mr. Mirci will also be on call during normal business hours of the second, third, and fourth phases of the project for any troubleshooting or support issues.

Mountain West Digital Library staff members, including director Sandra McIntyre, will provide feedback about the proposed data model, as well as assistance with distributing communications materials (detailed below) to MWDL partner institutions. As part of its regular activities, MWDL staff will re-harvest collections as the metadata in them is changed or updated as the result of project work. MWDL will also provide access to, and assistance with reviewing search statistics before and after the metadata values are changed.

As the primary partners on this project, metadata and cataloging librarians from **Brigham Young University and Utah State University** will provide consulting services and data for the following activities:

- Identifying collections that have a high percentage of local and regional names
- Providing lists of names from selected collections
- Exporting collections metadata at the beginning of the project, so the project has a baseline to measure changes
- Collaboratively review and assess possible data models to be used in project
- Collaboratively review and assess software packages to be used for expressing WNAF
- Review and test workflow for adding names to WNAF and provide suggestions for revision and enhancement
- Update local metadata collections with updated values based on WNAF.
- Participate in virtual meetings to collaborate on tasks throughout the project.

A recent survey of MWDL hubs has had responses from Southern Utah University, Utah State Archives, Utah State Library, Utah Department of Heritage and Arts, University of Oregon, Boise State University, and University of Nevada, Reno who are also interested in participating in the activities listed above. We anticipate that the time commitment for these additional partners will be lighter than for Utah State University and Brigham Young University.

An **Advisory Board** will be created to help guide this project through the four phases to make sure that the project goals are met and to provide input from different perspectives from national efforts working on similar efforts. The Advisory Board will consist of:

- **Philip E. Schreur** - Assistant University Librarian for Technical and Access Services, Stanford University's Green Library. Mr. Schreur has worked with the Linked Data for Libraries (LD4L) project and the upcoming Linked Data for Production (LD4P)[26] project.
- **Gretchen Gueguen** - Data Services Coordinator, DPLA
- A representative involved with BIBFRAME, potentially from Zepheira or a project utilizing BIBFRAME such as the University of California, Davis' BIBFLOW project
- A representative related to the Library of Congress Linked Data Service and/or the Program for Cooperative Cataloging

Co-PIs Jeremy Myntti and Anna Neatrour will be overseeing the project. The total cost for this two-year planning project is $50,000. Direct costs include: $12,270 to cover 5% of the Co-PI's time for 24 months based on their annual salaries; $9,120 to hire a student research assistant at a rate of $10/hr for 912/hrs; $4,819 for fringe benefits for the Co-PIs and the student; $6,970

[26] http://hangingtogether.org/?p=5195

for travel expenses; and $4,500 for services provided by partner institutions. Indirect costs include $12,321 to cover F&A budgeted at a rate of 32.7% in accordance with the University's federally negotiated rate for "Other Sponsored Activity."

5. **Toolkit and Communications plan**

This project will be developed in an open and accessible way, in order to provide the greatest amount of transparency and information for those participating in or interested in the project. Following general practices already established in the region for collaborative projects, we will create a Google site where we will develop areas for a toolkit, to cover the following:

- **Project introduction** -- Background information about the project, and a small selection of readings that can be used to place the project in context will be provided.
- **Data model assessment** -- Data models for a regional name vocabulary will be reviewed and assessed based on the needs of the University of Utah, partner institutions, and MWDL community. In order to accomplish this, a scoring model will be developed and data models will be collaboratively reviewed by the PIs, project partners, and MWDL staff.
- **Linked data vocabulary software evaluation** -- Open source software and other technologies that support the creation of LOD-ready controlled vocabularies will be reviewed and assessed, with a scoring model developed to assist with assessment. The template for the scoring model and the results of assessment will be shared.
- **Workflows and training materials** -- Workflows for reconciling and ingesting names into the selected solution for the pilot phase will be documented and tested at the University of Utah, Utah State University, Brigham Young University, and additional partners.
- **WNAF Dataset** -- The WNAF dataset will be available for download and reuse by other institutions. This will also provide examples of what the data used in this project looks like in a LOD-ready format.

In addition to the toolkit, we will communicate results at key project stages to the MWDL community through updates to community e-mail lists, blog posts, virtual meetings with MWDL staff and project partners, and webinars, as well as reporting out at regular MWDL Digitization Committee meetings. We anticipate providing an executive summary of the project upon conclusion with information on resources needed for larger scale implementation to the MWDL Governing Board and MWDL Digitization Committee. Results of this project will be shared with the library community through at least one scholarly publication and presented on at both regional and national conferences.

6. **Sustainability**

After this project has completed, we plan on developing a Project Grant proposal to expand the institutions contributing to the controlled vocabulary, offering training for those contributors, creating visualizations utilizing the relationship data in the controlled vocabulary, and developing a sustainable infrastructure to support the project.

Schedule of Completion - Year 1

	2016						2017					
	May	June	July	Aug	Sept	Oct	Nov	Dec	Jan	Feb	Mar	Apr
Activity	**Investigation Phase**						**Testing and Evaluation Phase**					
Kick off virtual meeting with project personel and partner institutions												
Advisory board meeting for investigation phase												
Collect and evaluate baseline data from fields with controlled vocabularies from partner institutions												
Capture baseline metadata from partner collections, and save for assessment activities												
Explore using EAC-CPF, OWL, and SKOS standards to create and represent relationships between entities within the controlled vocabulary												
Virtual meetings with partners to evaluate standards												
Summarize and communicate findings from first phase of work												
Advisory board meeting for testing and evaluation phase												
Develop a scoring model and evaluation criteria for open source software that supportes LOD vocabulary creation and maintenence												
Evaluate open source software that can be used to create, maintain, and make available the data contained in a compiled regional controlled vocabulary												
Virtual meetings with partners to evaluate software												
Develop plan for pilot and full evaluation of selected software												

Schedule of Completion - Year 2

	2017						2018					
	May	June	July	Aug	Sept	Oct	Nov	Dec	Jan	Feb	Mar	Apr
Activity	**Pilot Implementation**						**Assessment and planning**					
Advisory board meeting for pilot implementation phase												
Perform full evaluation of selected software - harvest, standardize, reconcile, and import controlled vocabulary information into the software of choice and make data available as LOD												
Hire and train student assistant to facilitate data entry, vocabulary reconciliation and enrichment, research, vocabulary maintentence, and assessment tasks.												
Enrich data with relationships and collections holding information.												
Explore the possibility of setting up an OpenRefine reconciliation service for the vocabulary												
Develop and revise collaborative workflows												
Virtual meetings with partners to revise workflows												
Advisory board meeting for assessment and planning phase												
Assess the outcomes of the project by reviewing workflow, identifying training opportunities, and exploring the impact of the centralized vocabulary on users of local digital asset management systems, regional digital library collaboratives (e.g., MWDL), and national level digital library efforts (e.g., DPLA).												
Capture and assess data on the percentage of names not in a national authority file, the number of names unique to one institution, and number of relationships we are able to express with the vocabulary.												
Develop a plan for expanding the controlled vocabulary to more institutions.												
Write executive summary of project for MWDL Board and Digitization committee												
Virtual meetings with partners for assessment and wrap-up												
Advisory board meeting for project wrap-up												
Present project at one or more national conferences												

Bibliography

Berners-Lee, T. (2006). Linked Data. Available at http://www.w3.org/DesignIssues/LinkedData.html

Dragon, P. (2009) Name Authority Control in Local Digitization Projects and the Eastern North Carolina Postcard Collection. Library Resources & Technical Services 53(3). doi:10.5860/lrts.53n3.185

Heath, T., and Bizer, C. (2011) Linked data: Evolving the web into a global data space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1. San Rafael, CA: Morgan & Claypool. Available at http://linkeddatabook.com/editions/1.0/

Ilik, V. (2015). Cataloger makeover: creating non-MARC name authorities. Cataloging & Classification Quarterly, 53(3-4), 382–398. doi:10.1080/01639374.2014.961626

IMLS Focus. The National Digital Platform for Libraries, Archives, and Museums. (2015) https://www.imls.gov/sites/default/files/publications/documents/2015imlsfocusndpreport.pdf

Myntti, J. and Neatrour, A. (forthcoming) Use existing data first: Reconcile metadata before creating new controlled vocabularies. Journal of Library Metadata, 15(3-4). doi:10.1080/19386389.2015.1099989

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

Veve, M. (2009) Supporting Name Authority Control in XML metadata: a practical approach at the University of Tennessee. Library Resources & Technical Services 53(1). doi:10.5860/lrts.53n3.185

Zeng, M. L., Lee, J., and Hayes, A. F. (2009) Metadata Decisions in Digital Library projects – Summary of an international survey. Journal of Library Metadata, 9(3-4). doi:10.1080/19386380903405074

DIGITAL STEWARDSHIP SUPPLEMENTARY INFORMATION FORM

Introduction
The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded research, data, software, and other digital products. The assets you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. However, applying these principles to the development and management of digital products is not always straightforward. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and best practices that could become quickly outdated. Instead, we ask that you answer a series of questions that address specific aspects of creating and managing digital assets. Your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions
If you propose to create any type of digital product as part of your project, complete this form. We define digital products very broadly. If you are developing anything through the use of information technology (e.g., digital collections, web resources, metadata, software, or data), you should complete this form.

Please indicate which of the following digital products you will create or collect during your project
(Check all that apply):

	Every proposal creating a digital product should complete	Part I
	If your project will create or collect	**Then you should complete**
☑	Digital content	Part II
☐	Software (systems, tools, apps, etc.)	Part III
☑	Dataset	Part IV

PART I.

A. Intellectual Property Rights and Permissions

We expect applicants to make federally funded work products widely available and usable through strategies such as publishing in open-access journals, depositing works in institutional or discipline-based repositories, and using non-restrictive licenses such as a Creative Commons license.

A.1 What will be the intellectual property status of the content, software, or datasets you intend to create? Who will hold the copyright? Will you assign a Creative Commons license (http://us.creativecommons.org) to the content? If so, which license will it be? If it is software, what open source license will you use (e.g., BSD, GNU, MIT)? Explain and justify your licensing selections.

This project will create an authority file of personal and corporate names. We do not assert any intellectual property rights over the data, and would dedicate it to the public domain with a CC0 license.

A.2 What ownership rights will your organization assert over the new digital content, software, or datasets and what conditions will you impose on access and use? Explain any terms of access and conditions of use, why they are justifiable, and how you will notify potential users about relevant terms or conditions.

We would not assert ownership rights over new digital content, software, or datasets. Narrative findings, workflows, and training materials developed as part of the project would be labeled with an Attribution-ShareAlike CC BY-SA license.

A.3 Will you create any content or products which may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities? If so, please describe the issues and how you plan to address them.

No, this project is going to deal with improving descriptive metadata for cultural heritage collections. There are no privacy or rights concerns for this work.

Part II: Projects Creating or Collecting Digital Content

A. Creating New Digital Content

A.1 Describe the digital content you will create and/or collect, the quantities of each type, and format you will use.

This project will consist of working with existing metadata and will not create or collect new digital content.

A.2 List the equipment, software, and supplies that you will use to create the content or the name of the service provider who will perform the work.

N/A

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to create, along with the relevant information on the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

We will not be scanning anything as part of this project. We will be creating a dataset in XML or RDF with an encoding scheme such as Encoded Archival Context - Corporate bodies, Persons, and Families (EAC-CPF), SKOS, or OWL.

B. Digital Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

Quality control will be conducted on the metadata collected in the project. Open Source tools such as OpenRefine will be used to sort and cluster data to ensure consistency of data. At least two people (e.g. student research assistant and one Co-PI) will review all metadata to make sure that errors are minimized.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance (e.g., storage systems, shared repositories, technical documentation, migration planning, commitment of organizational funding for these purposes). Please note: You may charge the Federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the Federal award. (See 2 CFR 200.461).

New digital assets will not be created, so there is no plan to preserve this type of data.

C. Metadata

C.1 Describe how you will produce metadata (e.g., technical, descriptive, administrative, or preservation). Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, or PREMIS) and metadata content (e.g., thesauri).

The project will create a database of metadata, but deciding on the standard to be used will be part of the research activity the project will undertake.

C.2 Explain your strategy for preserving and maintaining metadata created and/or collected during and after the award period of performance.

The Western Name Authority File (WNAF) database will be available on a server for download. We will also download and store the dataset locally in the University of Utah's Institutional Repository (USpace).

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of digital content created during your project (e.g., an API (Application Programming Interface), contributions to the Digital Public Library of America (DPLA) or other digital platform, or other support to allow batch queries and retrieval of metadata).

We will be developing and documenting workflows for partner institutions to use WNAF, and these will be available to the public. Metadata that is cleaned-up during the project in local systems will be harvested by the Mountain West Digital Library (MWDL) and Digital Public Library of America (DPLA).

D. Access and Use

D.1 Describe how you will make the digital content available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

WNAF will be openly available online. The main platform that will house the data will be determined as part of the project. The data will also be made available via the University's institutional repository.

D.2 Provide the name and URL(s) (Uniform Resource Locator) for any examples of previous digital collections or content your organization has created.

J. Willard Marriott Library Digital Collections: http://content.lib.utah.edu/
Western Soundscape Archive: http://westernsoundscape.org/
Mountain West Digital Library: http://mwdl.org/ (the Marriott Library provides hosting services for MWDL)

Part III. Projects Creating Software (systems, tools, apps, etc.)

A. General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) this software will serve.

A.2 List other existing software that wholly or partially perform the same functions, and explain how the tool or system you will create is different.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your software (systems, tools, apps, etc.) and explain why you chose them.

B.2 Describe how the intended software will extend or interoperate with other existing software.

B.3 Describe any underlying additional software or system dependencies necessary to run the new software you will create.

B.4 Describe the processes you will use for development documentation and for maintaining and updating technical documentation for users of the software.

B.5 Provide the name and URL(s) for examples of any previous software tools or systems your organization has created.

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under an open-source license to maximize access and promote reuse. What ownership rights will your organization assert over the software created, and what conditions will you impose on the access and use of this product? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain any prohibitive terms or conditions of use or access, explain why these terms or conditions are justifiable, and explain how you will notify potential users of the software or system.

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

C.3 Identify where you will be publicly depositing source code for the software developed:

Name of publicly accessible source code repository:
URL:

Part IV. Projects Creating a Dataset

1. Summarize the intended purpose of this data, the type of data to be collected or generated, the method for collection or generation, the approximate dates or frequency when the data will be generated or collected, and the intended use of the data collected.

Metadata will be collected from multiple partner institutions. This metadata will consist of Dublin Core metadata extracted from CONTENTdm, EAD finding aids, and other lists of terms currently residing in text files or spreadsheets. This data will be used to create a compiled names controlled vocabulary.

2. Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

No

3. Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

No

4. If you will collect additional documentation such as consent agreements along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

N/A

5. What will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

Metadata in text files, spreadsheets, and XML files will be collected and compiled into a single RDF data store.

6. What documentation (e.g., data documentation, codebooks, etc.) will you capture or create along with the dataset(s)? Where will the documentation be stored, and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

Documentation related to the source of data, data schema, alternate forms of names, related entries, and related collections will be stored on a Google site created for the project. This documentation will also be stored in text files along with the final dataset stored in the University's institutional repository.

7. What is the plan for archiving, managing, and disseminating data after the completion of the award-funded project?

The data created from this project will be the basis for expanding the controlled vocabulary to more institutions in the Mountain West. Because of this, the dataset will continually expand. A snapshot of the data will be created on a regular basis after the project has completed and archived in the University of Utah's digital preservation system.

8. Identify where you will be publicly depositing dataset(s):

Name of repository: USpace - the University of Utah's Institutional Repository
URL: http://uspace.utah.edu/

9. When and how frequently will you review this data management plan? How will the implementation be monitored?

This data management plan will be reviewed monthly throughout the project and quarterly after the two year project has completed.

Original Preliminary Proposal

Linking People: Developing Collaborative Regional Vocabularies

Across many institutions, controlled vocabularies for personal names and corporate bodies (hereafter names) are maintained in siloed information environments, for example as locally developed text fields within a particular CONTENTdm repository. Descriptive metadata work in the mountain west region could benefit greatly from a shared controlled vocabulary system for names, as librarians and archivists could draw on shared expertise about people or corporations that are notable regionally, but not likely to be within the scope of national vocabularies like the Library of Congress Name Authority File (NAF). Expressing controlled name vocabularies in a shared infrastructure that is Linked Open Data (LOD) compliant will help metadata catalogers in the region become more familiar with LOD technologies, as well as provide an infrastructure to visualize new connections between the entities represented in digital library collections.

This project draws on existing work of automating controlled vocabulary reconciliation that has been completed at the Marriott Library. Seventeen collections have had names matched with the NAF through a combination of vendor provided reconciliation and with scripts in OpenRefine. These collections have been updated with more accurate values, giving us a sample dataset of standardized local and regional names and an established workflow. These collections, along with additional data from partners, can form the basis of a LOD vocabulary for names in the region. This previous work has been shared at national conferences, and will be featured in an upcoming issue of the Journal of Library Metadata.

This planning grant will allow the Marriott Library to investigate, test, and pilot a workflow for developing the Mountain West Name Authority File (MWNAF), a shared vocabulary of names, which could later be expanded to include additional partner institutions of the Mountain West Digital Library (MWDL). The project will contribute to the National Digital Platform by providing a model that can be used by other partners of Digital Public Library of America (DPLA) service hubs as they move towards LOD and have the need to develop their own LOD compliant regional or local vocabularies. This project will provide a first step in standardizing names across MWDL partners, including libraries, museums, archives, and other cultural heritage institutions. Improving discoverability through the use of shared vocabularies will allow users to be more precise with their research within local, regional. and national discovery systems. Expressing the vocabulary as LOD provides the structure needed to make authority information open and repurposable.

It is estimated that the four phases of this two-year project will each take six months to complete. A detailed description of each phase is as follows:

1. **Investigation:** Collect and evaluate data from fields with controlled vocabularies from multiple partner institutions, such as the University of Utah, Utah State University, Brigham Young University, and other MWDL partners. Explore using the Encoded Archival Context - Corporate Bodies, Persons, and Families (EAC-CPF) standard to create and represent relationships between entities within the controlled vocabulary. Capture a baseline of analytics data by assessing how names in the selected vocabularies are currently discoverable in MWDL, DPLA, and Google. Adopt a data model for the vocabulary.
2. **Testing and Evaluation:** Collaboratively evaluate open source software that can be used to create, maintain, and make available the data contained in a compiled regional controlled vocabulary, with collaborative authority control. Develop a scoring model and

evaluation criteria that could be repurposed by other institutions with similar projects. Move forward with a pilot and full evaluation of selected software.

3. **Pilot Implementation:** Harvest, standardize, reconcile, and import controlled vocabulary information into the software of choice and make this data available as LOD. Enrich data with relationships and collections holding information. Explore the possibility of setting up an OpenRefine reconciliation service for the vocabulary. Document collaborative workflows and assess impact on work for the Marriott Library and partner institutions.
4. **Assessment:** Assess the outcomes of the project by reviewing workflow, identifying training opportunities, and exploring the impact of the centralized vocabulary on users of local digital asset management systems, regional digital library collaboratives (e.g., MWDL), and national level digital library efforts (e.g., DPLA). Other metrics that will be explored include capturing data on the percentage of names not in a national authority file, the number of names unique to one institution, and number of relationships we are able to express with the vocabulary. Develop a plan for expanding the controlled vocabulary to more institutions.

A related project is currently underway at the University of Nevada, Las Vegas (UNLV) with their Library Linked Data Project. While both institutions are working with personal name vocabularies, this project is centered on exploring workflows and software for multiple institutions to build and use collaboratively, and hopefully save both time and infrastructure costs. We plan to coordinate with UNLV to make sure the two projects complement each other.

The DPLA Metadata Application Profile (v.4), is architected to allow for harvesting LOD-ready Uniform Resource Identifiers (URIs) which will provide the ability to link directly to LOD triple stores such as the one that would be created in this project in the future. DPLA is currently harvesting URIs for spatial metadata from GeoNames with potential future plans to expand this service to other fields affected by local controlled vocabularies such as the one created in this project. The DPLA metadata team has been contacted in preparing this proposal and project personnel will continue to communicate with them throughout the project.

Results of this project will be shared with the library community through at least one scholarly publication and presented on the national level at both regional and national conferences. A toolkit providing an overview of the project along with forms and templates that can be repurposed will also be developed. After this project has completed, we plan on developing a Project Grant proposal to expand the institutions contributing to the controlled vocabulary, offering training for those contributors, creating visualizations utilizing the relationship data in the controlled vocabulary, and developing a sustainable infrastructure to support the project.

Co-PIs Jeremy Myntti, Librarian, Interim Head of Digital Library Services, and Anna Neatrour, Metadata Librarian, will be overseeing the project and working with partner institutions. The total cost for this project is $50,000. Direct costs include: $12,270 to cover 5% of the Co-PI's time for 24 months based on their annual salaries; $9,120 to hire a student research assistant at a rate of $10/hr for 912/hrs; $4,819 for fringe benefits for the Co-PIs and the student; $3,970 for travel expenses; and $7,500 for services provided by partner institutions. Indirect costs include $12,321 to cover F&A budgeted at a rate of 32.7% in accordance with the University's federally negotiated rate for "Other Sponsored Activity."