Implementing Name Authority Control into Institutional Repositories: A Staged Approach

Hannah Tarver, Laura Waugh, Mark Phillips, Will Hicks; University of North Texas Libraries; Denton, Texas, USA

Abstract

This paper outlines some issues with authority control in institutional repositories (IRs) and how the University of North Texas (UNT) Libraries have approached controlled vocabularies. In particular, it describes the staged approach used for implementing name authority in the UNT Digital Library, which has four steps: (1) put vocabularies on the Web as five-star open linked data, (2) make vocabularies available to metadata creators who actively use them, (3) store links, not strings, and (4) make data meaningful to users.

Background

Historically, libraries have focused not only on building collections, but on creating a means of describing the objects to make them findable and useful. Authority control structures function as a key component to usability by enabling optimal search and retrieval [5]. As the number of digital objects in library collections grows, organizations look to new approaches and system structures to support an increasing need for authority control mechanisms. Many organizations have realized the challenges of implementing an authority control structure; this issue remains an area of investigation and experimentation without a clearly defined solution [4].

Coupled with the innate difficulties of authority control in library collections, the increasing number of digital objects included in institutional repositories (IRs) offer unique challenges. The sheer increase of IRs and digital resources has produced a noted rise in the numbers of authors that must be included in name authority systems [3]. To compound this issue, users commonly want to search by personal names in IRs [6]; however, name variants create confusion, making it difficult to find all of the materials related to a particular person. In addition, some authors have the same name, causing additional difficulty in differentiation. Name authority control and management become even more complicated as IRs grow in size.

While large bodies of name authority records exist in a variety of formats -- such as campus-wide directories, local bibliographic catalog name authority records, the Library of Congress (LOC) Name Authority File,[1] the Virtual International Authority File (VIAF),[2] and even Wikipedia[3] -- these do not meet the needs of IRs for a variety of reasons. For example, a campus-wide directory may contain informal versions of faculty names. In addition, faculty often change institutions or publish with private sector partners not included in the directory [2]. Other projects focus only on current names or specific groups of people, such as The Names Project[4] for researchers in the United Kingdom (UK) or the International Standard name Identifier (ISNI)[5] for authors and artists.

Similarly, local bibliographic records, VIAF, the LOC Name Authority File, and Wikipedia are problematic because many authors do not have established names in these authority lists [2]. This was evident in a statistical sample of the University of North Texas (UNT) Libraries' IR name authority control database (see Table 1). In a random sampling of 100 authors in the UNT IR, only 26% are in the local bibliographic records, 32% have an authority file in the VIAF database, 28% have an authority file in the LOC Name Authority File, and 2% have a Wikipedia page. Additionally, most authors represented in one authority control file overlap with those represented in another; therefore the total number of authors represented in all of these authority files is still roughly one-third of those in the random sample.

UNT Digital Library Database	100 (random sampling)
Local bibliographic records	26
VIAF records	32
LOC records	28
Wikipedia	2

Table 1. VIAF, LOC, and Wikipedia Name Authority Record Comparison

To mitigate this growing problem, the UNT Libraries developed a locally-controlled authority structure – the UNT Name App[6] – as a step toward authority control in

[1] http://id.loc.gov/authorities/names.html
[2] http://viaf.org
[3] http://wikipedia.org/
[4] http://names.mimas.ac.uk
[5] http://www.isni.org/
[6] http://libdigital2test.library.unt.edu/name/

the UNT Digital Library. This endeavor is not an attempt to solve the larger problem, but to move forward and create a useful structure that can serve as a foundation within the local digital library system. The goals of the UNT Name App include: improving the quality of metadata records, streamlining the workflow for managing and maintaining these authority control lists, integrating these authority control lists into metadata creation workflows, and finally improving users' experience with the IR. Though the last goal is the most important, it is only obtainable if the other goals are met to provide the necessary infrastructure.

Implementation of Authorities

The approach at the UNT Libraries has its roots in the five stars of open data proposed by Tim Berners-Lee [1] which outline levels of accessibility to data that is open and reusable using existing technologies and mechanisms of the Web. The five stars are organized conditionally (each is dependent on meeting the previous requirements) and state that vocabularies/data must be available 1) on the web, 2) as a machine-readable structured data, 3) in a non-proprietary format, 4) using open standards from the World Wide Web Consortium (W3C), and 5) must link to other data to provide context.

Many successful initiatives provide a data dictionary and possibly an authorized list as a downloadable PDF; some also provide a machine-readable, structured version in HyperText Markup Language (HTML), Comma-Separated Values (CSV), or Microsoft Excel files. New initiatives such as Europeana and the Digital Public Library of America are immediately implementing all five of the stars as they move forward, creating positive momentum in this area that provides for more implementation possibilities.

The UNT Libraries development team began with five-star open linked data as the beginning point in the staged approach and outlined logical steps toward implementing name authority within the larger ecosystem of name authority control and open linked data. The four steps in this staged approach are:

1. Put vocabularies on the Web as five-star open linked data.
2. Make vocabularies available to metadata creators who actively use them.
3. Store links, not strings.
4. Make data meaningful to users.

These stages allow for necessary iterative changes to digital library infrastructure leading to improved findability and information collocation for users.

Simple Vocabularies

During the development of the content delivery system used by the UNT Digital Library, the implementation team decided that when a metadata field was limited to a controlled list of values, the terms should be publicly available in several open formats as a controlled list.[7] These lists were usable in a number of contexts in metadata creation, normalization, and viewing workflows. These simple vocabularies comprise a Django application for managing vocabularies and terms, and a set of views which expose these vocabularies in a variety of formats, such as Resource Description Framework (RDF), HTML, JavaScript Object Notation (JSON), and eXtensible Markup Language (XML) serializations.

Vocabularies handled in this way are basic lists that do not have a hierarchical structure or relationships, unlike name authority, which requires associated information for each name such as: alternate forms of the name, e-mail addresses or contact information, college/departmental or institutional affiliations, and relationships to other name authority vocabularies. However, the use of open linked data for simple vocabularies served as a testbed for more complex authority control at UNT.

Stage One: Building the App

The UNT Name App is a Django Web application that implements a set of common name authority operations. It includes the ability to distinguish between similar names with disambiguation strings, merge two or more records, and label variant names with a type, such as an abbreviation, translation, transliteration, or other format. Records can represent a person, organization, event, or software. Birth and death dates are stored separately in each record, and a free-text biography section utilizes Markdown[8] to provide a rich text area for information that does not fit into an existing field. There are a number of note field options including source, non-public, and deletion notes. Finally the model allows for links to other authority records with labels for common types, such as VIAF, LOC name authority file URLs, Wikipedia, Open Researcher and Contributor ID (ORCID),[9] and pages in the UNT Faculty Profile System.[10]

The UNT Name App provides data in several open, machine-readable formats including Metadata Authority Description Schema (MADS),[11] XML, and JSON. Additionally the HTML utilizes the Schema.org[12] vocabulary for names, providing another machine-readable format in Resource Description Framework in Attributes (RDFa). Each HTML page has embedded <link>

[7] http://digital2.library.unt.edu/vocabularies/
[8] http://daringfireball.net/projects/markdown/
[9] http://orcid.org
[10] http://faculty.unt.edu/
[11] http://www.loc.gov/standards/mads/
[12] http://schema.org

tags that reference other metadata representations of the record. The developers used appropriate HTTP response codes for features such as merged, suppressed, or deleted records. A simple Application Programming Interface (API) allows name lookups utilizing the Name Authority Cooperative Project (NACO) Normalization Rules and provides batch loading of names as well as programmatic lookup and resolution of names. Each name is assigned a unique number that is used to create the URL for the name, and which acts as the unique identifier for the record.

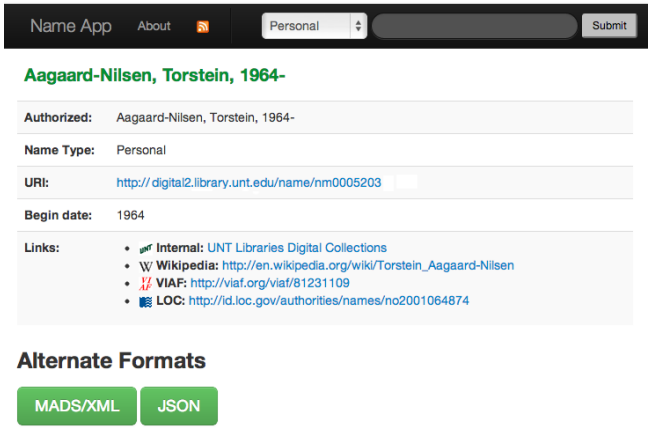

Figure 1: Example of an authority record in the UNT Name App.

Stage Two: Making it Available to Metadata Creators

Metadata creators and editors at UNT Libraries perform their normal operations within a form-based metadata editing application, also written in Django. The form includes helper utilities to facilitate the creation of normalized vocabularies and standards-based content. It provides autocomplete options for controlled vocabularies, drop-down list options for qualifiers, and validation against date/time standards,[13] as well as geo-spatial mapping tools.

Data-entry personnel have access to the data in the UNT Name App through a type-ahead feature on a select number of fields such as Creator, Contributor, and Publisher. As an editor types, Asynchronous JavaScript And XML (AJAX) requests are made to the UNT Name App on each keypress. The request searches across the fields and JSON responses populate a selectable list of candidate names that appears below the typed field. Because of the nearly-simultaneous interaction, editors can view approved names, disambiguations, and other information in real time (see Figure 2). At any point an editor may select a name and the typed information is replaced with the appropriate value from the UNT Name App.

[13] http://www.loc.gov/standards/datetime/

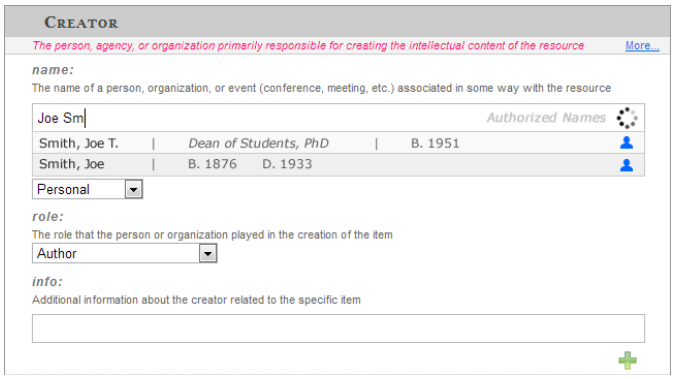

Figure 2. Example of type-ahead name tool in the metadata web form.

Individual names that appear within the type-ahead utility are linked to their respective records within the UNT Name App, giving editors quick access to enhance, expand, or alter the record information for the authorized name (e.g., to add new variants, connections, or other relevant data). Names can also be entered into a field manually if they are not yet in the Name App.

Stage Three: Storing Links

Once the information in the UNT Name App is available to metadata creators, the next step involves shifting the metadata coding from strings to links. Although humans can easily compare text strings, computers need something more concrete to distinguish a relationship between terms and entities.

For example, there are entries in the UNT system for Don Smith, Don W. Smith, Donald Smith, and Donald W. Smith. A computer may note that the level of similarity makes it probable that some (or all) of the names are the same person; however, it cannot make those final judgments. Some records with the names “Donald Smith” and “Donald W. Smith” are for items published by the National Advisory Committee for Aeronautics (NACA) during the 1940s-1950s and likely refer to the same person. Several other records include the name “Donald [W.] Smith” referring to a major professor or committee member for UNT theses and dissertations during the 2000s on topics in the biological sciences; that professor completed his bachelor’s degree in 1958, at the time of publication for many of the NACA reports. In this case, name authority should distinguish Donald Smith-the-aeronautics-engineer and Donald Smith-the-biology-professor as separate persons.

During this interim stage, nothing visibly changes for users but the infrastructure of the metadata and the system where it lives become more compatible with linked data (see Figure 3). Storing links which identify names or concepts allows for the possibility of using the data in more diverse ways at later stages in system development.

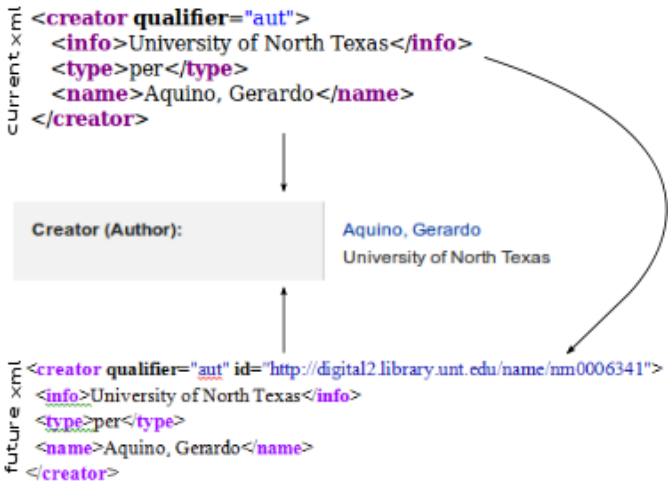

Figure 3: An example of metadata with name identifier links in the XML coding

Figure 3 contains an example (at the top) of the current XML coding used in the UNT Digital Library for a creator, in which the name and information/affiliation are stored as strings. Simple controlled vocabularies represent creator type (person) and role (author) and show up as codes in the XML. Beneath, a snippet of modified XML illustrates metadata with a unique URL identifier associated with that name stored in the coding. How the name displays to the user may not change -- shown in the center of the figure -- but the name could be displayed differently for different publications (e.g., as it appears on each item) while designating all records with the same URL name identifier as having the same creator or contributor.

Stage Four: Making Names Meaningful for Users

The most compelling examples of engaging user interfaces using name authority include projects such as the Open Libraries author pages[14] or WorldCat Identities.[15] In the commercial world, Amazon's[16] author and artist pages provide a similar experience, granting users a more precise discovery experience while augmenting it with linked and related data. These systems have an advantage over local IRs due to the vast amount of global information they can utilize to build these interfaces.

Although local IRs do not have the same massive amounts of data, they manage information and names that few (if any) of those larger entities control. Therefore, IRs should leverage the data that they have by creating link relationships from their local name authority files to other aggregations of names and information where appropriate. As IRs better understand how to expose name authority data to end users, these relationships will become increasingly important and will allow repositories to offer similarly-powerful services despite the smaller amounts of data available in IRs versus global systems.

[14] http://openlibrary.org/authors
[15] http://www.worldcat.org/identities/
[16] http://www.amazon.com

Steps for the Future

As libraries and digital collections grow and evolve, traditional methods of name authority and other forms of authority control must also change in order to stay relevant and offer the best possible services to users. Although the UNT Name App is currently a local solution, the iterative process that the UNT Libraries are using serves as one possible model for other institutions to implement name authority in IRs. Since this process is not tied to specific software, it allows for flexibility depending on available resources and software while following best practices for the Web. Improving name authority in local IRs also means the generation of more consistent, structured data and the potential in the future for a more powerful global network of repository data. Shifting authority control in this direction has the ability to improve findability and access for information professionals tasked with managing IRs and, most importantly, for all of the other users who want to discover those items.

References:

[1] Berners-Lee, Tim. "Linked Data." W3C. 7 July 2006. Web. 20 Feb. 2013 http://www.w3.org/DesignIssues/LinkedData.html

[2] Chapman, John W., Reynolds, David, and Shreeves, Sarah A. "Repository Metadata: Approaches and Challenges." Cataloging & Classification Quarterly 47:3/4 (2009): 309-325. Web. 18 Feb. 2013. https://www.ideals.illinois.edu/bitstream/handle/2142/13968/RepositoryMetadata_CCQ.pdf?sequence=2

[3] Hill, Amanda, Daniel Needham, and Alan Danskin. Names Project: Final Report. The Names Project, July 2009. Web. 20 Feb. 2013. http://www.jisc.ac.uk/media/documents/programmes/sharedservices/names-phase-one-final-report,.pdf

[4] Salo, Dorothea. "Name Authority Control in Institutional Repositories." Cataloging and Classification Quarterly 47.3/4 (2009). Minds @ UW. Web. 18 Feb. 2013. http://minds.wisconsin.edu/handle/1793/31735

[5] Summers, Ed. "Linking Things on the Web: A Pragmatic Examination of Linked Data for Libraries, Archives and Museums." arXiv. (2013). Web. 19 Feb. 2013. http://arxiv.org/ftp/arxiv/papers/1302/1302.4591.pdf

[6] Xia, Jingfeng. "Personal Name Identification in the Practice of Digital Repositories." Program: Electronic Library and Information Systems 40.3 (2006): 256-267. Web. 19 Feb. 2013. DOI: 10.1108/00330330610681330