10-2021

Name and Subject Heading Reconciliation to Linked Open Data Authorities using Virtual International Authority File and Library of Congress Linked Data Service APIs: A Case Study featuring Emblematica Online

Cindy Tang Tian

Name and Subject Heading Reconciliation to Linked Open Data Authorities using Virtual International Authority File and Library of Congress Linked Data Service APIs

A Case Study featuring Emblematica Online

Tang (Cindy) Tian, Timothy W. Cole, and Karen Yu

Libraries are actively exploring ways to use Linked Open Data (LOD) services to enhance discovery and facilitate the use of collections. Emblematica Online, *which provides integrated discovery of digitized emblem books, incorporates LOD in its design. As an implementation prerequisite, the Virtual International Authority File (VIAF) and Library of Congress (LC) Linked Data Service APIs were used to reconcile name and subject strings from legacy catalog records with global authoritative links from LOD resources. This case study reports on the automated reconciliation process used and examines the efficacy of the APIs in reconciling name and subject heading entities. While a majority of strings were successfully reconciled, analysis suggests that data cleanup, rigorously consistent formatting of metadata strings, and addressing challenges in existing LOD resources and services could improve results for this corpus.*

Tang (Cindy) Tian (ttian@nd.edu) is a Metadata Services Librarian at Kresge Law Library, University of Notre Dame. **Timothy W. Cole** (t-cole3@illinois.edu) (retired), Timothy W. Cole, Elaine and Allen Avner Professor Emeritus in Interdisciplinary Research, University Library, University of Illinois at Urbana-Champaign. **Karen Yu** (karen4@uchicago.edu) is Head of East Asia Technical Services, Joseph Regenstein Library, University of Chicago.

Manuscript submitted December 22, 2020; manuscript returned to authors for revision April 30, 2021; revised manuscript submitted June 14, 2021; accepted for publication July 2, 2021.

Emblematica Online is a web-based digital library that describes and supports the discovery of 1,406 retrospectively digitized facsimiles of rare emblem books that contain more than 33,000 individual emblems from seven research institutions: the Herzog August Bibliothek in Germany (466 books); the University of Illinois at Urbana-Champaign in Urbana, Illinois (421 books); the Getty Research Institute in Los Angeles (248 books); Duke University in Durham, North Carolina (197 books); Glasgow University in Scotland (43 books),

Utrecht University in the Netherlands (30 books); and the Newberry Library in Chicago (1 book). Early Modern emblem books expressed complex ideas in a compact and compelling form. Melding text and images, emblems (see figure 1) typically feature a tripartite structure consisting of a brief motto in Latin or a European vernacular language (*inscriptio*), an enigmatic illustration (*pictura*), and a textual epigram (*subscriptio*).[1] The emblem is more than the sum of its individual parts, however; the *inscriptio*, *pictura*, and *subscriptio* work together to produce a greater meaning, the goal of which is to challenge the reader intellectually and stimulate new thought and knowledge. Emblem collections were commonly published as books, but they also pervaded the decorative arts and appeared in other contexts. Analyses of emblems help scholars to develop a fuller understanding of both sacred and secular art of the period. The emblem is a critical genre in the study of Renaissance and Baroque Europe, owing both to its wide geographic spread and to the window it opens on the attitudes of the period concerning nearly every aspect of life, ranging from religion and politics to war and peace. Emblems suggest the presence of an intentioned, sophisticated strategy for repurposing, reorganizing, and reading texts and images through a system of parallels and analogies that narrow meaning to impart new perspectives or ideas.

Inherently, by their nature and because emblems embody both a rhetorical structure and a process, they are ideally suited to digital presentation in a Linked Open Data (LOD) context that can reflect semantic patterns of associative thought. For this corpus, a LOD approach enhances descriptive precision and facilitates interoperability across multiple, disparate, and widely distributed emblem book collections, thereby opening new ways for emblem scholars

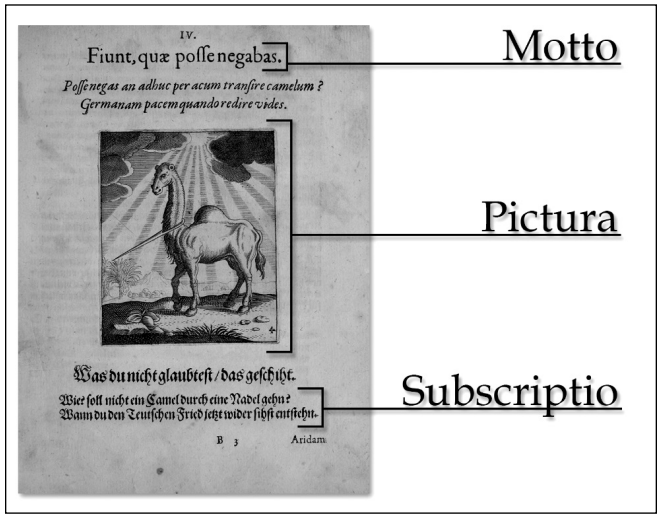

Figure 1. A tripartite emblem with inscriptio, pictura, and subscriptio.

to explore emblem literature. The LOD-based *Emblematica Online* portal makes emblems appearing in retrospectively digitized emblem books more visible to scholars in related disciplines, such as art historians, historians of Renaissance and Baroque cultures, comparative literary scholars, and other scholars who are interested in the wider relationship between literature and the visual arts, theories of representation, and iconography.

The original book-level and emblem-level metadata describing emblem book volumes and the individual emblems they contain were initially transformed by each participating library from local MARC records and local emblem-specific metadata records into records conforming to the Metadata Object Description Schema (MODS) and community-based emblem SPINE standard schemas, respectively.[2] Note that the development of the SPINE metadata structure standard was only one part of a larger effort toward a set of community metadata agreements for describing both emblem books and the individual emblems that they contain.[3] This work continues and to facilitate interoperability has included the adoption (guided by experience with the *Emblematica Online* portal and its precursors) of MODS usage guidelines and high level data content standards. To create the current incarnation of *Emblematica Online*, MODS and SPINE metadata records were harvested and normalized by scripts as needed. MODS/SPINE records are maintained in the portal backend as machine-readable XML.

LOD features and functionality have become an essential part of *Emblematica Online* to enhance discovery and research. The key point to enable these features is automated metadata reconciliation that maps bibliographic metadata from text strings to global Uniform Resource Identifiers (URIs) in LOD authorities (in this paper "LOD authorities" refers to LOD resources that can be used as substitutes for more traditional library authorities in the context of *Emblematica Online* and similar corpora). As part of the metadata reconciliation process for this project, a preexisting Python script for normalizing and managing MODS/SPINE metadata was adapted to integrate the reconciliation workflow and produce Resource Description Framework (RDF) graphs serialized as JavaScript Object Notation for Linked Data (JSON-LD), a way to store LOD in JSON format.[4] Names and subject headings in bibliographic records are two of the most representative metadata types that are suitable for exposure as LOD as there are more LOD on the web that provide contextual information around these classes of entities and include relevant relationships. An entity in the LOD sense of an entity-attribute-value model refers to who or what the authority value is about, as opposed to mere text strings in traditional authority control approaches.[5] Specifically, two tools are integrated in the script to query name and subject

heading entities respectively: the Virtual International Authority File (VIAF) Auto Suggest API (hereafter VIAF Auto Suggest API) and Library of Congress (LC) Linked Data Service APIs (hereafter LC Linked Data APIs).[6]

Hosted by OCLC, VIAF is a name authority service that coalesces authority files of different, mostly national, library institutions from around the world. Successful reconciliation of name entities with VIAF authority records can enhance the user experience of digital library collections by accessing new and analytic information such as name variations for an author, titles associated with the author, and name forms in different languages. VIAF's Auto Suggest API automatically searches authority terms within VIAF based on a text passed in a query. LC's Linked Data Service provides base Uniform Resource Locators (URLs) with various search constraints to query LC ontologies and controlled vocabularies. This project uses the *aLabel* search constraint that will "only return a resource whose authoritative label exactly matches the searched term."[7] The end goal of using these two APIs is to enrich the original SPINE and MODS XML metadata with VIAF and LC authoritative links for name and subject heading entities. For subject heading entities, in addition to topical subject headings, the queries also consider genre and geographic subject headings as subject heading entities to reconcile.

This case study examines the reconciliation process in particular with a focus on two key issues:

1. Understanding the efficacy of the APIs used to reconcile name and subject heading entities;
2. To identify solutions to improve match results of digital collection metadata reconciliation to LOD authorities.

Literature Review

Application of LOD in library contexts is an active, current area of research. The application of LOD features to library collections and resources both increases the visibility of these resources on the web and provides end users with enhanced representations of primary sources, search results, and analytic information for research, especially within digital library special collections.[8] As application of LOD gains momentum in libraries, it is important to recognize the essential role of metadata reconciliation as part of planning and implementing LOD within libraries. According to the five-star scheme for evaluating the quality of LOD implementations, an implementation reaches the five-star level when the entities mentioned in a web application's data and descriptions (expressed in accord with the RDF) are linked to other data sources and services on the Semantic Web.[9] For legacy data, e.g., bibliographic records describing emblem books, this is achieved by data reconciliation, which supplements names and subject headings with URLs linking to additional, authoritative information about these entities. Proper communication between original or legacy metadata and appropriate LOD authorities provides interoperability and standardization for existing collections, along with matching a greater percentage of terms to existing controlled vocabularies.[10]

Research done as part of the initial implementation of LOD features in *Emblematica Online* identifies a few of MARC's limitations for use with RDF, especially in contrast to MODS and other metadata schemas.[11] The same research shows preliminary statistical findings for transforming MARC string-based authority control terms into VIAF and LC Subject Headings (LCSH) links. Related research in the context of *Emblematica Online* also includes an analysis of the XML-based SPINE metadata schema and the transformation of corpus metadata to more RDF compatible ontologies.[12] The findings from this earlier research demonstrate that to facilitate discovery and enhance the value to scholars of digitized emblem books, metadata must first be enriched with additional URIs and the workflow upgraded to normalize and transform existing emblem metadata, recognizing that the effort to do this would be substantial and needed to be fully worked out.[13] Since this research was published in 2017, subsequent work has been done to create Python scripts to automatically identify select entities in legacy metadata that could be enriched with authoritative links to LOD resources. This study was motivated by a need to report on the automated reconciliation process and examine the match rates of a subset of entities to external LOD authorities using LOD services and associated search APIs.

Beyond the *Emblematica Online* corpus, other digital library collections have been used to experiment with reconciling authority headings against unique local thesauri using external tools such as OpenRefine.[14] These efforts include developing unique URL-generating applications in various formats for name entities, ongoing maintenance of local controlled vocabularies, and metadata reconciliation practices.[15] This has yielded positive results such as high match rates and personal name tracings not found in LC authority files and are "the first steps toward a more integrated conceptualization of authority work."[16]

While efforts have been made to create local controlled vocabularies to provide standardized terms for individual digital libraries, this approach is most "advantageous when digital collections use shared controlled vocabularies" or when objects in digital collections are unique to local institutions.[17] For entities with an existing authority record that was established following the standards organizations such as LC, linking to existing sources of controlled vocabularies provides the additional advantage of matching a greater

percentage of terms.[18] In a 2019 project to prototype and test data models for the LOD environment, the University of Maryland Libraries enhanced the local corporate name authorities by reconciling with the LC Name Authority File (LCNAF) where possible, taking advantage of the existing external data.[19] Beyond individual library collections, collaborative work across libraries is also being deployed to reconcile field objects against existing LOD authorities to prepare library data at large for a transition from MARC format to Linked Data. In the most recent Program for Cooperative Cataloging's (PCC) work on changes in MARC encoding to accommodate LOD identifiers, MARC field objects are reconciled to RDF URLs from VIAF and LCNAF as part of the process.[20]

Some literature emphasizes the importance of metadata clean up before the reconciliation. Van Hooland et al. state, "Before asking the question of how to link metadata from different sources, we need to develop strategies to check their initial quality and possibly solve issues that might disturb the reconciliation process among different resources."[21] Southwick indicated in a 2015 study on transforming digital collections metadata into LOD that the implementation process would be more efficient if metadata clean up was done to the extent possible before reconciling with LOD sets.[22]

Other suggestions from the professional literature recommend that existing LOD authorities and services also present challenges, and may not always be sufficient substitutes for more traditional library authorities. This issue was also of interest to the authors as they conducted their case study. In a 2013 study to determine which controlled vocabularies were best suited for use in a scientific data repository, White quoted the findings from 2007 preliminary research that there was not a single vocabulary adequate to describe an interdisciplinary field such as evolutionary biology.[23] The same gap existed, and still does, for libraries in general. A 2016 study by Radio and Hanrath addresses the issue of inadequate subject representation as affecting a resource's ability to interact with the LOD environment.[24] They call for increased attention and participation to identify areas of under- or misrepresentation in Linked Data vocabularies.[25] Whereas the current body of literature is focused on examining workflows and procedures for metadata reconciliation to LOD, there remains a need for more research examining, assessing, and reporting on the efficacy of the reconciliation services and tools used (as measured by final match result).

Method

This case study used a hybrid methodology that consisted of quantitative analysis and qualitative comparison to accomplish the name and subject heading reconciliation process for the *Emblematica Online* collection data. For the quantitative analysis, the authors dissected the XML MODS/SPINE bibliographic records and identified name and subject heading strings as entities for reconciliation. They retrieved a subset of items from the corpus and examined the efficacy of the VIAF Auto Suggest API and LC Linked Data APIs in matching the name and subject heading entities respectively.

VIAF and LC authorities have been extended to provide LOD services and APIs that increase the usefulness of these authorities. VIAF and LC authorities were selected as the LOD resources to which to reconcile name and subject heading entities in this case study due to the applicability of their scope and the extensiveness of Linked Data services they provide. The aforementioned preexisting Python script was adapted to generate statistics on the name or subject authoritative links to which each entity was matched in one query. The VIAF Auto Suggest API provides a fast lookup for authority records in VIAF and returns JSON blocks of personal or corporate name records with the *viafid* included as a unique identifier. Based on the granularity of the queried name string, the query can return one result, multiple results, or none. For example, when a name string lacks birth and/or death dates, the query can return multiple JSON blocks of different name authority records because these name entities cannot be disambiguated. For the purpose of accuracy, the Python script counts a match when only one *viafid* was found in the returned RDF (JSON serialization).

> VIAF Auto Suggest API: http://www.viaf.org/viaf/AutoSuggest?query=[query string]

For subject heading entities, this study identified multiple LC controlled vocabularies as the LOD authorities for different types of subject headings in the original metadata. These authorities included LCNAF for name subject headings, LCSH for topical subject headings, Library of Congress Genre/Form Terms (LCGFT) for genre subject headings, and MARC Geographic Areas (GAC) for geographic subject headings. Multiple base URLs/APIs were therefore constructed accordingly to reconcile the subject heading entities:

- LCSH search API for topical subject headings: http://id.loc.gov/search/?q=scheme:http://id.loc.gov/authorities/subjects&q=aLabel: "[query string]"
- LCGFT search API for genre subject headings: http://id.loc.gov/search/?q=scheme:http://id.loc.gov/authorities/genreForms&q=aLabel: "[query string]"
- LCNAF search API for name subject headings: http://id.loc.gov/search/?q=scheme:http://id.loc.gov

/authorities/names&q=aLabel:"[query string]"
- GAC search API for geographic subject headings: http://id.loc.gov/search/?q=scheme:http://id.loc.gov/vocabulary/geographicAreas&q=aLabel:"[query string]"

As previously noted, the *aLabel* search constraint only returns a result that exactly matches the term searched.[26] Therefore, the match number for each subject heading entity using the LC Linked Data APIs will be either zero (not matched) or one (matched).

In addition to a quantitative analysis, the study included a qualitative comparative analysis based on an interview with Deren Kudeki, HathiTrust Research Center (HTRC) Developer at University of Illinois at Urbana-Champaign who had done parallel work with HathiTrust catalog records, to better understand the use of these APIs in another institutional context. Following the examination and analyses of match rates of name and subject heading entities using the search APIs, this study intended to suggest (for subsequent research and confirmation) implementation techniques and solutions that improve reconciliation match results.

Reconciliation and Enrichment Using APIs

For *Emblematica Online*, book and emblem catalog records are stored in XML MODS/SPINE format and are freely retrievable across the web. To implement LOD, the name and subject heading entities in the original XML metadata were enriched with VIAF and LC authoritative links, and the XML was transformed (using XSLT) and saved as RDF (JSON-LD serialization). As noted, this was made possible by the integration of VIAF Auto Suggest API and LC Linked Data APIs. To examine the efficacy of these APIs as metadata reconciliation tools, a quantitative analysis was conducted by retrieving a subset of XML files from each of the six major institutions that participate in *Emblematica Online*. Since the University of Illinois at Urbana-Champaign (UIUC) and the Herzog August Bibliothek (HAB) hold most of the XML files (together 63 percent of the corpus), the study retrieved more files from these two institutions' collections than the other four. For name entities specifically, fifty metadata files (emblem books) each from UIUC and HAB collections and ten metadata files each from the Glasgow University, Utrecht University, Duke University, and the Getty Research Institute collections were randomly retrieved. For the subject heading entities, fifty metadata files from the UIUC collection and ten metadata files each from the Glasgow University, Duke University, and the Getty Research Institute collections were randomly retrieved. LCSH is not present in the original HAB or Utrecht University metadata files because of the limits on consensus to use LCSH by all partners from different nations, so subject metadata from these two institution collections was not used for the subject heading analysis.

The Python script incorporates *matchCount* as a new variable to track the number of matched authoritative link(s) by the VIAF Auto Suggest API or LC Linked Data APIs for a certain entity, and writes the results to a CSV file. When a name or subject heading entity is queried, the script first uses an *if* statement to check whether a VIAF or LC authoritative link (*valueURI*) is already present for that entity in the original XML metadata (i.e., previously reconciled). If a *valueURI* is present, the algorithm will skip that entity and move to query the next. This helps avoid skewed results regarding the efficacy of APIs by excluding entities that were previously reconciled. One exception is the name entities in the HAB collection. The majority of the name entities in the HAB collection have previously reconciled *valueURIs* that points to the Deutsche National Bibliothek (DNB) authorities. Since DNB is not within the scope of this study, the algorithm ignores DNB *valueURIs* and queries the name entities in the HAB collection using the VIAF Auto Suggest API. As mentioned, the script counts a match when only one result was returned (*matchCount* = 1).

End Results of Entity Match Counts

Name Entities

Table 1 shows the number of unique name entities in the retrieved metadata files from each institution collection. Table 2 summarizes the number of queried name entities, number of unique match counts, and calculates the match rates.

One thing to note is that the number of name entities that were actually queried ("Number Entities Queried" in table 2) equals the Unique Name Entities (in table 1) less the number of name entities that already have a *valueURI* in the original metadata file. This step is necessary to avoid skewed results. For example, the script found 118 unique name entities in the UIUC sample, among which 22 already have a *valueURI*. The algorithm skipped those 22 and queried the remaining 96 name entities, on which the calculation of match rate is based. However, this does not apply to the name entities from the HAB collection, since the algorithm was intentionally designed to query HAB name entities from a non-DNB name authority—VIAF. Therefore, the number of queried HAB name entities (267) remains the same.

As shown in table 2, only one name entity was queried and matched for the Utrecht sample. The sample size is

Table 1. Number of unique name entities

Collection	Files Processed	Unique Name Entities
HAB	50	267
UIUC	50	118
Duke University	10	12
Getty Research Institute	10	25
Glasgow University	10	13
Utrecht University	10	13
TOTAL	**140**	**448**

Table 2. Match rate of name entities using VIAF Auto Suggest API

Collection	Name Entities Queried	Match Count	Match Rate %
HAB	267	40	14.98
UIUC	96	60	62.50
Duke University	12	8	66.67
Getty Research Institute	18	9	50.00
Glasgow University	9	5	55.56
Utrecht University	1	1	100.00
TOTAL	**403**	**123**	**30.52**

Table 3. Number of subject heading entities

Collection	Files Processed	Unique Subject Heading Entities
HAB	0	0
UIUC	50	132
Duke University	10	13
Getty Research Institute	10	29
Glasgow University	10	13
Utrecht University	0	0
TOTAL	**80**	**187**

Table 4. Match rate of subject heading entities using LC Linked Data APIs

Collection	Subject Heading Entities Queried	Match Count	Match Rate %
HAB	0	0	N/A
UIUC	129	54	41.86
Duke University	13	7	53.85
Getty Research Institute	29	14	48.28
Glasgow University	13	10	76.92
Utrecht University	0	0	N/A
TOTAL	**184**	**85**	**46.20**

too small for the 100 percent match rate to be statistically meaningful. The match rate for the HAB sample (14.98 percent) is noticeably lower than others. One possible reason that may contribute to this low match rate is how many of the name entities in the HAB collection are formatted. They are formatted as name acronyms (which tend not to return matches) instead of full names. Another reason for match failure likely is that the lack of birth and/or death dates in many of the name entities returns too many results. As mentioned, match counts greater than 1 are not considered a successful match. Reason for the lack of dates in HAB name strings is unclear, but could be due in part to differences in metadata formatting and cataloging practices in Germany versus the US. As aforementioned, because of the differences in cataloging and metadata practices across nations (the partners participating in *Emblematica Online* span four countries), there were limitations on the guidelines that could be established for the participating partners to follow when creating the original metadata. That being said, the adoption of LOD gives potential for improving the consistency and richness of metadata with global authoritative links that provide end-users with disambiguated and enriched information. In this case study, because of the nuances in the original metadata, the match rate of name entities in the HAB sample is not very representative of the efficacy of the VIAF Auto Suggest API. Besides the highest match rate (100.00 percent for Utrecht) and the lowest (14.98 percent for HAB), the match rates of name entities using VIAF Auto Suggest API are between 50.00 percent-66.67 percent, with an average of 60.74 percent. With HAB included, the average match rate drops to 30.52 percent. This low HAB match rate suggests a need for further research as to why and to determine if a workaround is possible.

Subject Heading Entities

Table 3 presents the number of unique subject heading entities in the retrieved metadata files from each institution collection. Table 4 shows the number of queried subject heading entities, number of unique match counts, and calculates the match rates.

Similar to name entities and to avoid skewed results, the number of subject heading entities that were actually queried ("Subject Heading Entities Queried" in table 4) equals the Unique Subject Heading Entities in table 3 less the number of subject heading entities that already have a *valueURI* in the original metadata file. Since LCSH is not present in the original HAB and Utrecht University metadata files, no subject metadata from these two institution

collections was used for the subject heading analysis. Table 4 shows that the match rates of subject heading entities using LC Linked Data APIs are between 41.86 percent-76.92 percent, with an average of 46.20 percent.

HathiTrust Research Center (HTRC) LOD Project

As a reality check and to better appreciate the facets of this corpus that might influence reconciliation, the authors compared their results to those found for the HathiTrust Research Center (HTRC) LOD Project. The HTRC project works with metadata describing 17 million volumes across different institution's libraries to create BIBFRAME records from MARC records. The project reconciles contributor names to VIAF and subject headings to LC authorities. The same open source APIs and services are used—VIAF Auto Suggest API for name entities and LC Linked Data APIs for subject heading entities. Within the scope of LC controlled vocabularies, GAC is searched for geographic subject heading entities and LCSH for other subject headings. In 2019, name entities of over 17 million HTRC volumes had a match rate of 75.00 percent from VIAF, and subject heading entities of the same corpus had a match rate of 15.00 percent.[27] HTRC's match rate of 75.00 percent for its name entities is higher than the average match rate of name entities in this case study (60.74 percent not including HAB match rate, 30.52 percent including HAB match rate). This high match rate of the HTRC project was achieved by "using different ways to finesse the queries such as getting rid of the parentheses, and trying both a full date and just the start year in date."[28] The match rate for subject heading entities of the HTRC project (15.00 percent) is lower than the average match rate of subject heading entities in this case study (46.20 percent). Based on the interview and the authors' observations, they extrapolate some of the explanations for this difference:

- As a specialized collection, the subject headings in the *Emblematica Online* corpus are more uniform, such as "*Emblems*," "*Conduct of life*," "*Love in art*," etc. that already have an established heading in the LC authorities. Subject headings in the HTRC corpus, in contrast, are much broader, with more than 17 million volumes on various subjects. It is possible that LC's Linked Data APIs respond better to specialized collections in reconciling subject heading entities, but more work is needed to prove that point.
- The HTRC project reconciled its general subject heading entities to LCSH and geographic subject heading entities to GAC. By contrast, *Emblematica Online* expanded to include LCGFT, LCNAF, and MARC Countries as part of the LOD authorities used for the reconciliation in addition to LCSH and GAC. The use of multiple LOD authorities improved the match rate by matching genre and name subject heading entities to authoritative links that do not exist in LCSH or GAC.

Discussion

To transform digital library collection metadata into Linked Data, it is essential to implement a successful reconciliation that finds the best match to authoritative links for name and subject entities. Lessons learned from and the challenges during the reconciliation process of this case study are discussed below.

Prep Work before Reconciliation

It is important to minimize the metadata errors in the original metadata files. For example, one name entity in the HAB collection "*a. b. c. d. e. f. g. h. i. k. l. m. n. o. p. q. r. s. t. u. w. x. y. z.*" was erroneously recorded and returned no match result in VIAF. The correct form of the name in VIAF is "*A a b c d e f g h i k l m n o p q r r s s t u w x y z.*"[29] Several letters ("A," "r," and "s") were missing in the original name string. As a result, the authoritative link was not found by VIAF's Auto Suggest API. It is also important to ensure that the data is formatted correctly during processing, especially for non-English texts that involve diacritics. For example, one name entity "*Mabre Cramoisy, Sébastien*" was stored as "*Mabre Cramoisy, Sâebastien*" in the original XML metadata, which returned no match in VIAF because the Unicode character "*é*" was mistakenly transformed to "*âe*" during the data ingestion from the institution to *Emblematica Online*.

Data heterogeneity remains a challenge for metadata cleanup. It is hard to maintain consistency for heterogeneous digital collections when metadata is integrated from different sources or various data providers, as is the case for *Emblematica Online*. Van Hooland et al. pointed out that metadata quality and inconsistency will continue to remain a challenge for the reconciliation to LOD due to a lack of established methodologies or tools for metadata quality evaluation.[30] Specific to this case study, more consistent and standardized metadata would also have required more manual work on the legacy metadata and reaching consensus about matters of practice that have long varied across national boundaries. Metadata errors and incorrect ingested data in this case study were greatly minimized by the long-standing collaborations among the partners that led to the adoption of the SPINE schema, MODS usage guidelines, and high-level data content standards. Even so, as described above, enough variability in metadata

remained to create some challenges that interfere with the reconciliation process using the same API.

During Reconciliation

Discovering techniques to manipulate and format data strings is often needed to improve the match rate. During the reconciliation process of *Emblematica Online*, the authors experimented with two techniques to prepare metadata in a way that was proven to help find a unique match in VIAF:

- Changing angle brackets to square brackets. For example, no match was returned when using VIAF's Auto Suggest API for the name entity "*Sibylla Ursula <Braunschweig-Lüneburg, Herzogin>*" that was in the original metadata, but a match was found when the angle brackets were changed to square brackets and querying "*Sibylla Ursula [Braunschweig-Lüneburg, Herzogin]*".
- Removing punctuation at the end of a name string. For example, a unique result was returned for the name entity "*Mello, G. de*" but not for "*Mello, G. de.*".

However, it is worth noting that these formatting techniques vary by name and are difficult to anticipate in code. Depending on the LOD authorities and how the entity is formatted in that authority, one technique typically cannot apply to all entities across diverse collections (i.e., with metadata from diverse sources). For example, the name entity "*Josephus <Romanorum, Rex, I.>*" does not have a match in VIAF with either angle or square brackets. Similarly, "*Mauclerc, Antonius.*" returns a unique match result regardless of whether the period is present at the end of the string. The inconsistency of these formatting techniques presents challenges in preparing original or legacy metadata for reconciliation because there is no single solution to various formatting issues. As a result, it is up to the libraries and LOD practitioners to discover and implement what works best for their collection data.

LOD Resources as Authorities

It might be every LOD practitioner's dream that a single LOD authority contains all quality authority records that can be easily reconciled to by various entities. When White quoted the preliminary research conducted in 2007 that no single vocabulary was adequate for describing an interdisciplinary field, it was not clear that the same issue would be exemplified in today's ever-growing LOD implementation attempts.[31] For example, in LC Linked Data Service, geographic names are established in LCNAF, GAC, and MARC Countries, but not in LCSH.[32] This means that to automate reconciliation of a geographic name entity used as a subject heading, LOD practitioners need to query other controlled vocabularies different from LCSH, such as LCNAF or GAC, to find a match to the authoritative link in LC Linked Data Service. By contrast, in traditional authority control practice, geographic names that can be assigned as geographic subject divisions can be easily searched manually by librarians in both "Name Authority Headings" and "Subject Authority Headings" using the LC authorities interface.[33] The ambiguity and inconsistency in how LOD resources connect to traditional library authorities like the LC authorities presents a challenge, and raises the question of whether LOD resources can be considered as encompassing the function and role of traditional library authorities.

Conclusion

This study describes the reconciliation of name and subject heading entities of *Emblematica Online* and examines the efficacy of the VIAF Auto Suggest API and LC Linked Data APIs in reconciling metadata to LOD authorities. Results from the quantitative analysis indicate that the average match rate of name entities using VIAF Auto Suggest API is 60.74 percent (without HAB match rate), and 30.52 percent (with HAB match rate). The average match rate of subject heading entities using LC Linked Data APIs is 46.20 percent. This study identifies solutions to improve match results of the metadata reconciliation in three aspects—data cleanup, formatting metadata strings, and paying attention to the ambiguity and inconsistency in how LOD resources connect to traditional library authorities.

The authors' case study adds to the growing body of work examining the application of LOD best practices to library special collections. The findings on the efficacy of VIAF Auto Suggest API and LC Linked Data APIs and the lessons learned through the course of this work can potentially be useful to personnel managing other digital libraries who are contemplating similar LOD reconciliation projects. Implementation tools and techniques in this study are easy to use and could provide opportunities for the larger digital library community to engage in incorporating LOD into the catalog.

However, the corpus used in this case study is limited to one specialized digital collection and only a small portion of the total corpus data was examined. A subsequent phase of research should extend the approach used here to the records of the entire corpus, refining the current approach to enhance the reconciliation match results. One possible direction for increased experimentation on this corpus would be to compare the scope and coverage of different LOD resources such as Wikidata, the Getty Art and

Architecture Thesaurus (AAT), the Bibliothèque nationale de France (BnF authority file), etc. Also, although the current approach yielded good reconciliation results for most institution collections in the *Emblematica Online* corpus, they did not work well for certain institutions. For example, the match result for name entities in the HAB collections using the VIAF Auto Suggest API was significantly lower than that of the other institution collections. The reasons for this need to be investigated further in a subsequent phase of work. This paper speculated the possible reasons based on observations, but it also shows the need to investigate the systematic disparity among different institution collections that would affect the final reconciliation results.

References and Notes

1. *The Princeton Encyclopedia of Poetry and Poetics*, 4th ed. (Princeton, NJ: Princeton University Press, 2012), s.v. "EMBLEM."
2. "Metadata Object Description Schema," Library of Congress, last modified February 5, 2020, http://www.loc.gov/standards/mods/; Stephen Rawles, "A Spine of Information Headings for Emblem-Related Electronic Resources," in *Digital Collections and the Management of Knowledge: Renaissance Emblem Literature as a Case Study for the Digitization of Rare Texts and Images*, ed. Mara R. Wade (Salzburg: DigiCULT Project, 2004), 19–28, https://www.digicult.info/pages/special.php.
3. Nuala Koetter, "Interoperability of Digital Emblematica Metadata Using the Open Archives Initiative Metadata Harvesting Protocol and Other Schemas," in *Digital Collections and the Management of Knowledge: Renaissance Emblem Literature as a Case Study for the Digitization of Rare Texts and Images*, ed. Mara R. Wade (Salzburg: DigiCULT Project, 2004), 79–87, https://www.digicult.info/pages/special.php; Thomas Stäcker, "Transporting Emblem Metadata with OAI," ed. Mara R. Wade (Salzburg: DigiCULT Project, 2004), 89–95, https://www.digicult.info/pages/special.php.
4. "EmblematicaOnline," accessed May 28, 2021, https://github.com/cindyttian/EmblematicaOnline/blob/main/end2end-rev5Dec2019.py.
5. *Linked Data Glossary*, s.v. "Entity," by W3C Working Group Note, last modified June 27, 2013, https://www.w3.org/TR/ld-glossary.
6. "VIAF—Authority Cluster Auto Suggest," OCLC Developer Network, accessed December 2, 2020, https://platform.worldcat.org/api-explorer/apis/VIAF/AuthorityCluster/AutoSuggest; "Technical Center," Library of Congress Linked Data Service, accessed December 2, 2020, https://id.loc.gov/techcenter/searching.html.
7. "VIAF—Authority Cluster Auto Suggest."
8. Katrina Fenlon et al., "Exploring Linked Data Benefits for Digital Library Users," *Proceedings of the Association for Information Science & Technology* 55, no. 1 (2018): 799–800, https://doi.org/10.1002/pra2.2018.14505501122.
9. Tim Berners-Lee, "Linked Data," July 27, 2006, https://www.w3.org/DesignIssues/LinkedData.html.
10. Jeremy Myntti and Anna Neatrour, "Use Existing Data First: Reconcile Metadata before Creating New Controlled Vocabularies," *Journal of Library Metadata* 15, no. 3–4 (2015): 205, https://doi.org/10.1080/19386389.2015.1099989.
11. Timothy W. Cole et al., "Library MARC Records into Linked Open Data: Challenges and Opportunities," *Journal of Library Metadata* 13, no. 2–3 (2013): 163–96, https://doi.org/10.1080/19386389.2013.826074.
12. Timothy W. Cole et al., "Using Linked Open Data to Enhance the Discoverability, Functionality and Impact of Emblematica Online," *Library Hi Tech* 35, no. 1 (2017): 159–78, https://doi.org/10.1108/LHT-11-2016-0126.
13. Cole et al., "Using Linked Open Data."
14. Scott Carlson and Amber Seely, "Using OpenRefine's Reconciliation to Validate Local Authority Headings," *Cataloging & Classification Quarterly* 55, no. 1 (2017): 1–11, https://doi.org/10.1080/01639374.2016.1245693; Silvia B. Southwick, "A Guide for Transforming Digital Collections Metadata into Linked Data Using Open Source Technologies," *Journal of Library Metadata* 15, no. 1 (2015): 1–35, https://doi.org/10.1080/19386389.2015.1007009.
15. "django-name," accessed October 19, 2020, https://github.com/unt-libraries/django-name; Jeannette Ho, "Name Disambiguation for Digital Collections: Planning a Linked Data App for Authority Control at Texas A&M University Libraries" (presentation at the LD4 Conference on Linked Data in Libraries, Boston, MA, May 11, 2019).
16. Carlson and Seely, "Using OpenRefine's Reconciliation," 10.
17. Southwick, "A Guide for Transforming Digital Collections Metadata," 20.
18. Myntti and Neatrour, "Use Existing Data First," 191–207.
19. Bria Parker and Adam Gray, "Rethinking the University of Maryland Authority File for the LOD Environment," *Journal of Library Metadata* 19, no. 1–2 (2019): 69–81, https://doi.org/10.1080/19386389.2019.1589699.
20. Jackie Shieh, "PCC's Work on URIs in MARC," *Cataloging & Classification Quarterly* 58, no. 3–4 (2020): 418–27, https://doi.org/10.1080/01639374.2019.1705951.
21. Seth van Hooland et al., "Evaluating the Success of Vocabulary Reconciliation for Cultural Heritage Collections," *Journal of the American Society for*

Information Science and Technology 64, no. 3 (2013): 469, https://doi.org/10.1002/asi.22763.
22. Southwick, "A Guide for Transforming Digital Collections Metadata," 19.
23. Hollie White, "Examining Scientific Vocabulary: Mapping Controlled Vocabularies with Free Text Keywords," *Cataloging & Classification Quarterly* 51, no. 6 (2013): 655–74, https://doi.org/10.1080/01639374.2013.777004.
24. Erik Radio and Scott Hanrath, "Measuring the Impact and Effectiveness of Transitioning to a Linked Data Vocabulary," *Journal of Library Metadata* 16, no. 2 (2016): 80–94, https://doi.org/10.1080/19386389.2016.1215734.
25. Radio and Hanrath, "Measuring the Impact and Effectiveness of Transitioning to a Linked Data Vocabulary," 92.
26. "Technical Center," Library of Congress Linked Data Service, accessed December 2, 2020, https://id.loc.gov/techcenter/searching.html.
27. Deren Kudeki (HathiTrust Research Center Developer at University of Illinois at Urbana-Champaign), interview with author, November 15, 2019.
28. Kudeki interview with author, November 15, 2019.
29. "A a b c d e f g h i k l m n o p q r r s s t u w x y z," VIAF, accessed October 11, 2020, http://viaf.org/viaf/33061207.
30. Van Hooland et al., "Evaluating the Success of Vocabulary Reconciliation for Cultural Heritage Collections," 469.
31. White, "Examining Scientific Vocabulary," 658.
32. "Library of Congress Names," Library of Congress Linked Data Service, accessed June 8, 2021, https://id.loc.gov/authorities/names.html; "MARC List for Geographic Areas," Library of Congress Linked Data Service, accessed June 8, 2021, https://id.loc.gov/vocabulary/geographicAreas.html; "Library of Congress Subject Headings," Library of Congress Linked Data Service, accessed June 8, 2021, https://id.loc.gov/authorities/subjects.html.
33. Library of Congress Authorities, accessed June 8, 2021, https://authorities.loc.gov/.