National Strategy for Shareable Local Name Authorities National Forum : White Paper

2018-03-29

Principal Investigators: Chew Chiat Naun and Jason Kovari

Authors (alphabetical order): Michele Casalini, Chew Chiat Naun, Chad Cluff, Michelle Durocher, Steven Folsom, Paul Frank, Janifer Gatenby, Jean Godby, Jason Kovari, Nancy Lorimer, Clifford Lynch, Peter Murray, Jeremy Myntti, Anna Neatrour, Cory Nimer, Suzanne Pilsk, Daniel Pitti, Isabel Quintana, Jing Wang, Simeon Warner

This project was made possible in part by the Institute of Museum and Library Services [LG-73-16-0040-16]. The views, findings, conclusions or recommendations expressed in this publication do not necessarily represent those of the Institute of Museum and Library Services.

Table of Contents

1. Background

Funded by the Institute for Museum and Library Services (IMLS), the National Strategy for Shareable Local Name Authorities National Forum (SLNA-NF) brought together stakeholders working on local name authorities in order to develop a shared understanding of key issues facilitating and preventing sharing of these data. SLNA-NF aimed to explore a range of possible avenues and outcomes to further the objective of making authorities more shareable.

The April 2015 IMLS Focus Summary Report: National Digital Platform[1] emphasized the importance of enabling technologies (e.g., interoperability via linked data) and radical collaborations in supporting the mission of the cultural heritage sector. The SLNA-NF fits well within this context: colleagues gathering to consider deeply the issues surrounding local authority practice in an attempt to identify enabling technology, infrastructure, data models, social contracts, collaborative approaches and more. As such, we sought to include stakeholders representing a wide range of disciplinary and institutional perspectives.

As a first step, it was necessary to survey the variety of use cases and assumptions in each domain, and to understand preconditions for improving shareability. To do so, we held a series of virtual discussions during which the following points emerged as essential to examine: workflows, data models, persistence, institutional mandates, and communities.

Following five months of virtual meetings, the group held its first in-person meeting at Cornell University (October 2016) where themes focused on workflow, service models, responsibility and ethical issues, tooling and infrastructure, best practices, and community. A summary of that meeting is available on the SLNA-NF wiki: https://confluence.cornell.edu/x/bKpRF

Participants met virtually on these topics and reconvened for a second in-person meeting at the Library of Congress (April 2017). During this second meeting, we focused our efforts on examining practices from existing institutions, collaborations and tooling; further, we began diving into the issues at-hand for the areas examined in the following report: minimum viable specifications, data provider obligations, workflows and reconciliation as a service. A summary of the meeting's proceedings is available on the SLNA-NF wiki: https://confluence.cornell.edu/x/BuF0F

Related resources:

- Project Wiki: https://confluence.cornell.edu/x/inf_Ew
- IMLS Proposal: https://www.imls.gov/grants/awarded/lg-73-16-0040-16

[1] IMLS (2015). *IMLS Focus Summary Report: National Digital Platform*: https://www.imls.gov/publications/imls-focus-summary-report-national-digital-platform

2. Project Participants

Micah Altman (MIT)
Michele Casalini (Casalini Libri)
John Chapman (OCLC)
Robert Chavez (NEJM)
Chew Chiat Naun (Harvard University)
Chad Cluff (Backstage Library Works)
Joan Cobb (Getty Research Institute)
Mike Davidson (National Library of Medicine)
Michelle Durocher (Harvard University)
Nancy Fallgren (National Library of Medicine)
Steven Folsom (Cornell University)
Paul Frank (Library of Congress)
Janifer Gatenby (OCLC Leiden)
John Gieschen (Symplectic Elements)
Jean Godby (OCLC)
Eric Hanson (Johns Hopkins University)
Corey Harper (Elsevier Labs)
Kirk Hess (Library of Congress)
Timothy Hill (Europeana)
Diane Hillmann (Metadata Management Associates, LLC)
Kate James (Library of Congress)
Tom Johnson (DPLA)
Jason Kovari (Cornell University)
Dean B. Krafft (Cornell University)
Nettie Lagace (NISO)
Ted Lawless (ThomsonReuters)
Nancy Lorimer (Stanford University)
Clifford Lynch (CNI)
Sally McCallum (Library of Congress)
Andrew MacEwan (British Library)
Worthy Martin (University of Virginia)
Peter Murray (IndexData)
Jeremy Myntti (University of Utah)
Anna Neatrour (University of Utah)
Corey Nimer (Brigham Young University)
Suzanne Pilsk (Smithsonian Institution)
Daniel Pitti (University of Virginia)
Tiziana Possemato (@Cult, Casalini Libri)
Isabel Quintana (Harvard University)
Carl G. Stahmer (UC Davis)
Hannah Tarver (University of North Texas)
Jing Wang (Johns Hopkins University)
Simeon Warner (Cornell University)
Ryan Wick (Oregon State University)

Principal Investigators: Chew Chiat Naun (04/16-04/17); Jason Kovari (04/17-04/18)

3. Introduction

Libraries, government agencies and other cultural heritage and educational institutions create name registries, agent entities, or authorities to serve a variety of purposes, usually within an institutional or disciplinary context; meanwhile, data these organizations create have potential for broader reuse. In considering sharing of local name authorities, issues arise regarding workflows, data models and persistence.

Workflows vary widely. Library authority practice is based on creation of data by third party producers (e.g., a cataloger), but researcher identifier systems typically rely on self-registration (e.g., ORCID). Creation of data may be performed within a local environment or on a a highly controlled central platform shared by many contributors (e.g., the Name Authority Cooperative Program or NACO). Ingest or production processes may be set up to pre-empt duplication, or be relatively tolerant of it. If the latter, downstream reconciliation becomes more important for data consumers. Data from any of these sources may be aggregated into a central hub (e.g., ISNI); there may or may not be a quality threshold for publication of the data. Sharing of data raises the issue of propagating changes and keeping data in sync. As always, there will be trade-offs between workflows that are considered optimal and what the available infrastructure will support.

Data models need to be addressed to facilitate sharing. The vocabularies used to express these relationships also need to align if data are to be reused effectively across domains. The need to support reuse in itself introduces the need for new relationships, such as explicit declarations of equivalence or of non-equivalence. This is an area where linked data offers advantages; however, reconceiving legacy practices in linked data terms can pose challenges that different communities are at varying stages of addressing. These issues arise not only for newly created data, but even more acutely for the large bodies of legacy data still in need of reconciliation. Further, divergence from strict adherence to shared data models due to historical contingencies (e.g.: staffing, local interests, organizational structure and/or legacy systems) pose obstacles to shareability. Even when acknowledging these limitations, practical considerations may mandate that our adopted models accommodate legacy systems in the near-term. In addition, communities will have to navigate the policy issues that arise in a linked environment, such as the use of alternative vocabularies; prioritization between widely-established; heavily-linked identities and unique, local, or siloed ones; or issues related to the contribution threshold.

Persistence of both the identifier and the resource's data is an issue that arises repeatedly in discussions about reuse. Best practices often include non-reassignment of identifiers, tombstoning of deprecated identifiers, and retention of selected data. These prescriptions are readily implemented in some areas, such as traditional library authorities, but present a greater hurdle to accustomed practices in other contexts. In an ecosystem where data are shared and modified by multiple entities, provenance and trust loom as fundamental issues; this raises questions about data modeling, identifier management and data management, ownership and data

quality control policies. Many of these issues around persistent identifiers have been assessed in by the Australian National Data Service.[2]

Differing institutional mandates may be reflected in the implementation of divergent practices. Organizations as diverse as national libraries, universities, scholarly publishers, value added service providers, and disciplinary societies will have different focuses; thus, resulting use cases will show different emphases. These differences will make themselves felt in crucial areas such as the business models that sustain their services or sources of funding. For example, a university may provide support for a researcher identifier system to track the scholarly output of currently affiliated faculty, but it is less clear if the same justification extends to past scholars or to the broader scholarly community. Yet shared data must be able to meet a range of use cases outside its original purpose.

The issues outlined above will appear throughout the following sections of this paper. Through the lenses of minimum viable specification, data provider obligations, workflows and reconciliation services, the SLNA-NF explored key issues that may either facilitate or hinder sharing of locally produced and stored data.

4. Minimum Viable Specification

4.1. Introduction

This section considers how a Minimum Viable Product (MVP) approach may be applied to identity data. The goal of this section is not to propose a MVP specification, but to consider how an Agile development[3] approach should influence identity management practice; in other words, a minimally viable product upon which continual iterations facilitate the team's learning process and system improvements. More specifically, this section will focus on data requirements, technical requirements, and business requirements. Finally, essential questions such as those concerning licensing of data will be raised.

The term "minimum viable product" has acquired a pejorative connotation in some quarters, perhaps because of a perception that it encourages poor quality metadata. This perception confuses the process with the outcome. Wikipedia describes MVP in these terms: "a minimum viable product (MVP) is a product with just enough features to satisfy early customers, and to provide feedback for future product development."[4] This approach is well suited to identity data, which is often created with limited (which is not always to say negligible) context but lends itself to verification - or falsification - and enrichment as it is subsequently reused. The key word here is "viable": what set of functional requirements does the minimum specification need to support?

The use cases that received most attention in SLNA-NF revolved around aggregation of identity data. While not all communities place a premium on sharing of data, it was a value held in

[2] Australian National Data Service (n.d.). *Persistent Identifiers: Expert Level,* retrieved 2018-03-27 http://www.ands.org.au/guides/persistent-identifiers-expert#sec321
[3] *Agile software development*, retrieved 2018-03-25. https://en.wikipedia.org/wiki/Agile_software_development
[4] *Minimum Viable Product*, retrieved 2018-03-25. https://en.wikipedia.org/wiki/Minimum_viable_product

common among SLNA-NF participants. In the current environment, sharing has been achieved through large-scale centralized aggregations such as VIAF (Virtual International Authority File) and ISNI (International Standard Name Identifier). Sharing can also serve to spotlight issues where authorities data could be improved, for example in the case of Digital Public Library of America's (DPLA) large scale metadata aggregation efforts, where names in digital collections are often inconsistent.

While local authorities designed for use cases not intended for data aggregation exist, these datasets were out-of-scope for SLNA-NF, which placed sharing as essential for consideration.

4.2. Data Requirements

In the shared environment model relevant to SLNA-NF participants, there must be minimum data requirements in order to support machine matching and disambiguation at scale. Since experience strongly suggests that human intervention is essential to effective management of these data, the requirements must support human interpretation as well; as one example, the Social Networks and Archival Context (SNAC) project has developed methods for human intervention in disambiguation of entities represented in their prototype tool.

Merging unique identifiers is an important issue to consider as information about entities becomes more robust through collaborative maintenance and research activities. The ISNI data quality policy[5] contains provisions that could be said to model the MVP approach. These provisions include requirements that entries cannot be added to the ISNI database without basic elements such as name and an identifier from the source agency. The policy goes on to outline a scoring system for adequacy of data.

The ISNI data quality policy stops short of prescribing best practice to contributing communities. This is a further step that some communities are now contemplating. The Program for Cooperative Cataloging (PCC) Task Group on Identity Management in NACO[6], for example, is considering an approach where some combination of attributes (e.g., name + affiliation + work) would be considered adequate to establish an identity, while other attributes fulfill a corroborating role.

In SLNA-NF discussions on this question, two points gained wide acceptance. The first is that attributes supporting disambiguation are often highly domain-specific. Although name + date, facilitates many disambiguation tasks, disambiguation attributes must focus on domain-specific data when this is combination is not sufficient; for example, instrument is a strong differentiator for musicians whereas more familiar attributes like affiliation are less relevant.

The second is that it pays to be pragmatic about the choice of potential differentiating attributes captured during the metadata creation process. Provenance metadata captured in the course of

[5] ISNI (n.d.). *Data Quality Policy*. http://www.isni.org/content/data-quality-policy
[6] Program for Cooperative Cataloging (2016). *Charge for PCC Task Group on Identity Management in NACO*. https://www.loc.gov/aba/pcc/documents/Identity-management-NACO-PCC-TG.pdf

metadata interventions can provide a strong trail of evidence to confirm or contradict hunches that human operators can have about identity. For example, a simple name entry for "David Smith" would not help users to collocate or disambiguate; however, if we knew this entry was supplied by the "GPTD, Guild of Professional Teachers of Dance," the provenance of the data helps us place the entity within the world of professional dance. Similarly, some substantive information is easily recorded in certain kinds of projects; e.g., affiliation is usually known when an institution registers its own faculty. This type of activity can be extended beyond the initial metadata creation stage to be iterated throughout the metadata lifecycle; an institution querying a registry to identify film directors, for example, could enhance the description with that information when a positive identification is made.

The idea of "designing for shareability" may become increasingly influential as it gains acceptance. Other practices from the linked data domain may also gain traction, such as leveraging existing domain identifiers to confirm identifications (e.g., IMDb for the above mentioned film directors example). Certain operating assumptions that come from a closed system may need reappraisal in an open world environment[7]. For instance, it cannot be presumed that two identifiers do not refer to the same entity under an open world assumption. Thus, metadata created for cross-domain reuse will need to be more explicit in making assertions about identity and non-identity.

In the past, the focus of identity management has often been on properly identifying an entity by ascribing a work or other creative outputs to that entity. A secondary focus has been on clearly stating that this entity was definitively different from other entities within a set of metadata, as a precondition for supporting further assertions about that entity. The indexing of library authority data in search engines coupled with our desire to facilitate reuse of our entities in new contexts (e.g., Wikipedia and Wikidata) bring this secondary focus to the forefront. Furthermore, this further sharing facilitates authors and other stakeholders to stumble across their library-created identities, e.g., through a simple Google search; as such, some authors have contacted metadata creators to align their works with their personal entities. Making clear assertions that an entity is unique within a set of metadata - and across other datasets - is an important function, even when we know little else about the entity.

From the viewpoint of the data consumer, persistent identifiers are a functional requirement that are not essential when working entirely in a closed system; however, these persistent identifiers are crucial in an open world environment. Thus, a strong business need exists for two thresholds of viability: one for internal management purposes and the other for general publication. While it may be expedient - and highly efficacious for metadata management purposes - to register entities without concern for a persistent identifier (e.g., provisional status with fragmentary data), this approach creates problems for data consumers; provisional identifiers may change and are often not dependable links to the described entity: the latter being the motivation of using identifiers in end-user facing applications. The notion of provisionality is recognized by existing identity management organizations, including ISNI where it exists in distinction to an "assigned" status. Provisional is a distinction that also exists in NACO, where these are defined as

[7] Sequeda, J. (2012). *Introduction to: Open World Assumption vs Closed World Assumption*. Dataversity. http://www.dataversity.net/introduction-to-open-world-assumption-vs-closed-world-assumption/

"Authority records with Level of establishment 008/33 value c. These may be created by NACO participants when there is not enough information to establish a full record. Examples: The contributing library does not have the language expertise to establish the record in full; or a subordinate body name is available only in a language that is different from the language of the parent body."[8].

In linked data practice there is a strong presumption that a persistent URI is the preferred way of referencing entities and accessing facts about them. Yet not only in legacy data, but also in the real world, information may be available in less tractable forms. A question that will only be answered through experience is the benefit of capturing uncontrolled data, or in minting entities on the fly. It seems likely that the MVP approach will need to accommodate a variety, or a hierarchy, of techniques for capturing what is known about an entity. Under this model, versioning best practices and metadata for versioning (e.g. deprecated entities and their replacements) will need consideration.

For the process of iterative corroboration and enhancement of metadata to take place effectively, another requirement has to be met: comparing data from different sources. It will be difficult to match on organization affiliation, for example, unless the data can be reliably compared; this includes not only the identity of the organization, but the relationship between the person and that organization. Further, organizational entities themselves present another potential alignment issue in that these often have imprecise boundaries, may be described by different communities with different levels of granularity, or are lack clarity concerning the organization's substructures.

Dates represent another area where alignment of data is more difficult when the models and levels of granularity differ; date information can be difficult to compare if the dates are for different things. In one example cited by an SLNA-NF participant, a date representing a period when an individual was professionally active was inadvertently mapped to the individual's birth date, causing automated matching to fail, thus producing duplicate entries for the individual in an aggregator. Work on mapping, aligning attributes, and relationships among sources has to go hand-in-hand with work on determining minimum viable product specifications.

4.3. Technical Requirements

Technical requirements, including infrastructure, are central to the notion of sharing. Systems intended for sharing must provide robust query and disambiguation services, often in conjunction with frequently updated data dumps. Systems must be able to manage distribution of updates to various partners. As use of these services increases, scalability becomes critical. Scaling is important for both data storage as well as the services that make accessible the dataset. In many ways, this is a precondition for identifier persistence in that persistence is not possible without infrastructures capable of supporting the amount of data produced or the frequency of requests for serving that data; if not scaleable, the system will be unusable.

[8] Program for Cooperative Cataloging (2005). *NACO Participants' Manual, 3rd Edition*. https://www.loc.gov/aba/pcc/naco/documents/npm3rd.pdf

Available tools often reflect the needs of the purpose they originally served and the workflows they were embedded in; however, tools need to be evaluated against a broader range of uses to meet evolving needs. One of the clearer areas of need is the ability for vocabularies to integrate with external lookups to editing tools and reconciliation services. Look-up services greatly facilitate the use of diverse datasets for metadata practitioners. Effective user-interfaces provide quick responses during query and browse, and this necessitates centralized indexes over the data. In the Linked Data for Libraries - Labs project, work is underway to extend the Samvera community's Questioning Authorities gem to provide contextual look-up for a variety of linked data sources (e.g.: id.loc.gov, Getty Vocabularies, FAST, etc.)[9]; this is designed to be implementable across repository architectures.

The SLNA-NF believes that further consideration of technical requirements in an MVP context is an important research area for future work.

4.4. Business Requirements

Any approach, including MVP, must consider business models; these include considerations around funding, data licensing and governance. While business models can be lightweight, sharing of data is not a sustainable practice for any dataset without consideration of these issues.

Generally, libraries preference unrestricted data licenses in hopes of facilitating as much reuse of data as possible. While not completely analogous in that this refers to descriptive metadata about resources rather than data about identities, the DPLA Metadata Application Profile includes principle regarding restricting data:

> The DPLA believes that the vast majority of metadata as defined herein is not subject to copyright protection because it either expresses only objective facts (which are not original) or constitutes expression so limited by the number of ways the underlying ideas can be expressed that such expression has merged with those ideas. To be protectable, a work must be original, which means that it must contain at least a "modicum" of creativity in its creation, selection, or arrangement. Facts and ideas may not be copyrighted. Even if a work is original, it may be limited by the doctrine of "merger," which states that when there are a limited number of ways an idea can be expressed, the idea merges with the expression, and is therefore not subject to copyright. These two limitations on the application of copyright are the reason the vast majority of metadata is not subject to copyright protection.[10]

Licensing of metadata is a difficult issue for related reasons. While unrestricted licensing of data is often cited as an ideal, it may be in conflict with the business model that supports the service

[9] Linked Data for Libraries - Labs (2016-2018). *Samvera (aka Hydra) Community Linked Data Support.* https://wiki.duraspace.org/x/84E2BQ

[10] Digital Public Library of America (n.d.). *Metadata Application Profile*, accessed 2018-03-28. https://pro.dp.la/hubs/metadata-application-profile

in the first place, and some compromise may need to be struck. For example, Ringgold supplies a subset of its metadata to ISNI, all of which, except the Ringgold ID is available via the ISNI public interfaces. Ringgold withholds richer data and the ID that will access it from ISNI to protect its value-added services. Other contributors of data to the ISNI database restrict the display and distribution of their data for either commercial or legal reasons. The restrictions apply to all but the ISNI Quality Team; where data, such as titles of associated works, are restricted, it can be detrimental because as it can prevent merging and splitting decisions by ISNI members and registration agencies. As partially demonstrated with the above ISNI examples, conflicts of interest arise when reusing data in an open world environment; this is particularly relevant where business models and profit expectations misalign between stakeholders.

Further, concerns around confidentiality poses additional issues when considering identity data aggregation and reuse. While much library-originated identity data is unlikely to raise alarms around personal information, legislative considerations need to be addressed as part of the formation of a business model. For instance, the European Union's General Data Protection Regulation (GDPR) will begin being enforced as of May 2018; this effort is "designed to harmonize data privacy laws across Europe, to protect and empower all EU citizens data privacy and to reshape the way organizations across the region approach data privacy."[11]

Among considerations in the GDPR legislation are the right of EU citizens to learn what private information stakeholders have about them as well as the right of EU citizens to be forgotten; in the EU context, the right to be forgotten pertains to an individual's being able to prevent "being perpetually or periodically stigmatized as a consequence of a specific action performed in the past, especially when these events occurred many years ago and do not have any relationship with the contemporary context."[12] In response to this concept, IFLA produces a statement on the right to be forgotten, which includes issues for libraries; notably, IFLA's statement points to the need to balance preservation of the historical record with patrons' privacy concerns that underlie this right.[13] Again, library data is unlikely to hold sensitive information about individuals; however, legislative issues, such as those arising the GDPR, should be considered as organizations build and aggregate identity data.

Finally, communities must address governance to respond effectively to needs ranging from compatibility of data models to scalable infrastructure to sustainable business models. Alongside governance structures for their data and underlying services, communities need to clearly articulate to their audience their goals as well as their place in landscape in which they operate. Because these are areas of shared need and, at times, conflict of interests, they are potential areas for collaboration and negotiation to achieve mutual benefits for the community.

The SLNA-NF believes that further consideration of business requirements in an MVP context is and important research area for future work.

---

[11] *European Union General Data Protection Regulation Portal*, accessed 2018-03-28. https://www.eugdpr.org/
[12] Mantelero, A. (2013). *The EU Proposal for a General Data Protection Regulation and the roots of the 'right to be forgotten'*. Computer Law & Security Review (29,5), p. 230. http://dx.doi.org/10.1016/j.clsr.2013.03.010
[13] IFLA (2016). *IFLA Statement on the Right to be Forgotten*, accessed 2018-03-28. https://www.ifla.org/publications/node/10320

5. Data Provider Obligations

5.1. Introduction

This goal of this section is to develop a framework that identifies the obligations of providers of library authority data. We consider the special requirements of library authorities and similarly curated inventories of people, places, organizations and events such as Wikidata and Geonames. Working on the front lines in organizations that manage significant aggregations of library authority data, we offer our own perspectives, expressed as recommendations, or "lessons learned", which may or may not have been implemented. At the end of the discussion, we use the framework to assess the compliance of a few important authority-data aggregations.

A useful starting place is the ontology published by W3C's Data on the Web Best Practices Working Group.[14] At the top of the ontology is a list of 14 challenges[15], reproduced in Table 1 below. A click on one of the headings reveals the requirements that address the challenge, as well as links to use cases collected by the W3C that illustrate their application. For example, the "Identification" challenge is met when each resource is associated with a unique identifier, and the "Vocabularies" challenge is addressed when term lists or ontologies are documented, openly shared, and reused. As one of the use cases shows, a dataset that conforms to the Linked Data conventions by Tim Berners-Lee extends one of these requirements by specifying that identifiers must also be persistent.

Table 1. Challenges identified in W3C's Best Practices for Data on the Web ontology.

Access	Usage	Enrichment
Vocabularies	Formats	Licenses
Granularity	Metadata	Identification
Preservation	Quality	Provenance
Selection	Sensitive Data	

Using VIAF as an example of a library-community data resource, it is easy to demonstrate the value of the W3C framework as a checklist for the assessment of data provider obligations. For example, VIAF might earn high marks on the "Identifier" challenge because each page has a unique URI that is reasonably persistent. But the "Vocabularies" challenge identifies room for improvement. VIAF's XML format is an adaptation of the MARC Authority standard that accommodates clusters of headings, a primary data structure. But since no description of this specialized version of the MARC Authority standard is publicly accessible, VIAF's consumers must infer the semantics of the affected fields. Several other challenges listed in the W3C

[14] W3C (2015-02-24). *Data on the Web Best Practices Use Cases & Requirements*. W3C Working Group Note. https://www.w3.org/TR/2015/NOTE-dwbp-ucr-20150224/

[15] W3C (2015-02-24). *Data on the Web Best Practices Use Cases & Requirements. 3.2 Requirements by Challenge*. W3C Working Group Note. https://www.w3.org/TR/2015/NOTE-dwbp-ucr-20150224/#requirements-by-challenge

framework are also relevant to VIAF, such as Sensitive Data, Provenance, and Enrichment; these points are discussed in more detail later in this document.

This discussion make a distinction between the roles of data creator/publisher of an individual authority file such as LCNAF, and a data aggregator such as VIAF; this distinction is made because the two types of datasets present different challenges. For example, data creators/publishers must take special care to represent uniquely identifying information, while data aggregators must track the provenance of the data that comprises their collection. Aggregators must also highlight the enrichments that make the whole greater than the sum of its parts. This said, services can function as both an aggregator as well as a creator/publisher.

The rest of the section on Data Provider Obligations is organized into sections. Section 5.2 identifies special characteristics of authority data that cannot be addressed in detail by the W3C framework. Section 5.3 explores two possible extensions to the W3C framework: data quality and data synchronization. Section 5.4 considers how the challenges of data quality, sharing, diffusion, scalability, and synchronization are addressed in ISNI, NACO, VIAF, and SNAC, with a more detailed set of recommendations tailored to the needs of library authority data.

5.2. Unique characteristics of authority data

The W3C recommendations, especially those regarding access, licenses, formats and vocabularies are largely implemented in the library and archive communities. Yet, following these recommendations alone will not achieve an optimal result due to the unique characteristics of authority metadata.

Foremost, there are many specific aspects of quality that are important to authority control that may not manifest in other types of data aggregation, notably data that facilitates disambiguation. Additionally, many workflows demand synchronization, which is not explicit in the W3C recommendations. The below section concentrates on these specific data and environmental characteristics, and the measures necessary to handle data correctly.

5.2.1. The authority-control environment.

The main purpose of library-based authority control is to accurately relate publicly available cultural and intellectual works to the persons and organizations involved in their creation. To do so, authority data are created with just the right amount of information to distinguish among identities with the same name without impinging on privacy. These data are then used to create links to existing works and performances, and later to link to future works and performances, as well as to previously undiscovered or unlinked works. Thus, linking authority data to works of all resource types and performances (and vice versa) is an ongoing process.

Authority data with links to creative works facilitate searching, enabling more precise and complete discovery. They are also used for collecting a complete set of works to aid in reputation management, as in grant applications, university ratings, and fundraising. They may

also be used in rights management, where the unambiguous links between works and creators or contributors are essential in correctly directing payment.

Authority data shared with data aggregations and distributed as linked data promise to expose significant parts of a library collection and connect it in a web of relationships to related people, places, organizations, concepts, and other creative works.

5.2.2. Achieving scale.

The horizons of authority data are widening to be cross-purpose and cross-domain, and to include more comprehensive links to a wider range of publications, curated databases and data stores in addition to resources in traditional library collections. Because creating authority records and validating links between authority data are time consuming tasks, an increasing number of players must collaborate to achieve the scale required by these newer, broader goals. Interoperating with publishers, rights-holding systems, internet search engines and commercial aggregators is complicated by conflicting business and technical models, which may obscure common interests and the benefits of scale in working together. These differing approaches manifest throughout our processes, one example being the treatment of pseudonyms. Wikipedia collects all pseudonyms together with a person's real name. Libraries vary; some treat pseudonyms as separate identities and some as name variants. Finally, Rights Management societies almost always treat pseudonyms as separate identities.

Redundancy, and steps to avoid redundant data, is another aspect to achieving scale. Authority files that are machine-generated from bibliographic data in library catalogs are likely to produce lower-quality data than identities hand-crafted; successfully deduplicating these identities among themselves and alongside other sources in an automated process is difficult or impossible, which will cause problems when merging with higher-quality data. When such data are ingested into an aggregate file, further amassing, merging and sorting of data occurs, with only a relatively small promise of success. This speaks to a need for aggregators to track data provenance and build logic for data confidence into aggregation algorithms to avoid incorrect merges and duplication. Without this, scale may be achieved but quality will not.

5.2.3. Sharing.

Authority data are shared in various ways. They may be created directly in a multi-source database (e.g., NACO or ISNI) and thus immediately shared with other contributing sources. Creating directly in a larger file allows the data creator to disambiguate the subject-identity against a larger pool of data. Authority data created in smaller databases (e.g., local authorities) may be made available in downloadable files or as copies of individual records available via searching or linking; further, these data may be somehow aggregated into the aggregator / multi-source authority files. A completely decentralized model with a huge network of links and copied data makes both the diffusion of corrected data and the task of disambiguation extremely difficult. Thus centralization is desirable, arguably essential. Centralization may be achieved either through centralized creation and maintenance, or through controlled aggregation and disambiguation of data from multiple sources.

Authority data have characteristics that are significantly different from other data types, and these differences make it difficult to apply models of sharing and distribution that are already applied to bibliographic records and documents. For example, authority data raise the following issues:

- Name and identity ambiguity
- Pseudonyms and hidden identities
- Mixed identities, or single entities that describe multiple, unique entities
- Frequent updates
    - for individuals, such as changing affiliations and life-event dates
    - for organizations, such as changing organizational structures
- Extensive "sameAs" links, accurate or otherwise
- Privacy requirements

Thus, when making a list of recommendations for data creators and aggregators, the recommendations from W3C[16] and the linked data community[17] can serve as generic guidelines but require additional precision for capturing the unique characteristics of authority data.

5.3. Challenges with authority data

W3C recommendations on quality are difficult to apply to authority metadata (e.g., requirements relating to data sets and completeness); however, many specific recommendations exist for authority metadata.

Regarding distribution of data, the W3C recommendations are explicit and generally well implemented when distributing from aggregations such as VIAF, NACO, ISNI, SNAC, and the SHARE Virtual Discovery Environment (Share-VDE). Data formats use published standards such as MARC 21, EAC-CPF and ISNI and these are distributed in syntaxes including ISO 2709 (MARC), RDF/XML, JSON and n-triples. Published data vocabularies are widely used such as MARC vocabularies on the Library of Congress web site.

5.3.1. Quality.

It is necessary to find the right balance between machine processing and manual effort to achieve acceptable quality at-scale. Quality control for authority data has two agents: the data creator and the data aggregator. Defining quality for both agents through metadata application profiles is critical and requires constant negotiation. Data creators are responsible for the quality of the data within the collection over which they have editorial control; data aggregators strive to maximise the value of existing authority data by matching, consolidating, normalizing, and adding other

[16] W3C (2015) *Data on the Web Best Practices Use Cases & Requirements*. W3C Working Group Note 24 February 2015, Section 4. https://www.w3.org/TR/dwbp-ucr/#requirements-1
[17] W3C (2014). *Best Practices for Publishing Linked Data.* W3C Working Group Note 09 January 2014. https://www.w3.org/TR/ld-bp/

valuable enhancements. Both data creators and data aggregators must take efforts to avoid introducing problems such as duplicative data and mixed identities.

It would be desirable for providers and aggregators to convey a sense of confidence in their data. Confidence is not easy to convey; however, some practices already implemented can be interpreted as measures of quality. For example, it is possible to indicate the level of certainty about an identity's date. Further, inclusion of URIs to sources of information is a good indicator of high confidence in the metadata. In standard formats such as MARC 21, there is no particular place, except in notes fields, to indicate whether the metadata has been derived directly, e.g. by phone call to an author or a publisher or from local institution records. It would be helpful to be able to distinguish data derived from bibliographic sources from that derived from more direct sources. For example, ISNI gives more weight to information coming from rights management societies, that have direct contact with authors, than that coming from trade sources, that are more likely deriving data from published materials.

Aggregated sources present additional data quality issues. For example, when data without widely recognized disambiguating identifiers (e.g., NACO, VIAF, ISNI and ORCID) are included in other files, there is a reliance on algorithms to match and merge. These processes can introduce errors. For example, machine algorithms merge data concerning the same identity while keeping different identities apart, but they are not 100% effective when data coming from different sources varies in the perceived or measurable levels of quality, or in the detail of the information available. When source records lack sufficient information for matching, they cannot be confidently merged, generating either a mixed identity or a duplicate, the latter being a less serious error. An aggregated file requires a means to correct errors in a timely fashion and to effectively diffuse the corrections.

These data quality issues are not solely derived from aggregation; the same issues may derive from data creators; different workflows employed by data creators can create records that are not significantly similar to records made by different creators for the same identity in the "eyes" of matching and merging algorithms. For example, some data creators may create the data manually with careful cross-checking, while others may generate authority data algorithmically from bibliographic records, meaning that there will be both some percentage of duplicates and some percentage of mixed identities.

5.3.1.1. Data quality and automatic matching and merging

Data that are most useful when shared beyond their source of creation must be disambiguated in a broader context. As the environment of authority data expands, disambiguation needs to be international and cross-domain in scope.

Aggregation combined with a disambiguation process is also called *reconciliation*, which can be implemented with automated or manual procedures. Their effects are complementary. The automated process reconciles by creating links that assert "this is the same identity" but they are not verified as true. In contrast, such links that are created manually are verified. The ISNI Quality Team provides a "post-aggregation" manual verification service by examining and

responding to end user input and also by periodic sampling to pinpoint areas of the database that would most benefit from manual verification and correction, where necessary. Similar manual reconciliation functions are provided in other services, as well. For example, the SNAC project treats identity records for personal names as identity clusters in their aggregated environment of authority records. A new identity cluster created by a user will trigger a reconciliation process, and a manual review of potential duplicates before a new cluster is created. Identity clusters are also sent through a process of review in the SNAC system, ensuring that quality control takes place in the aggregated environment.

5.3.1.2. Automated error correction

It is possible to devise tests to detect data errors and sometimes correct them by algorithm. Examples of such tests employed by VIAF and ISNI include date anomalies, such as persons publishing before the age of nine or dying before their birth date. In the ISNI database, the birth date of 1 January in a given year is the most common birth date because of a default that is applied in some databases when only the year is known. These dates are regarded as suspect and will not prevent a merge from occurring where there is a match on the year. The date anomaly program catches these errors, but such tests need extensive testing before they can be deployed. Detecting mixed identities is a bigger challenge. Where incoming authority data matches more than one record and matching multiple sources exist on both or all potential match candidates, the incoming record can be flagged as suspect.

Similarly, some aggregations have programs to normalize data from different sources. For example, ISNI has a suite of programs that aim to treat pseudonyms consistently as related names, not name variants.

5.3.1.3 Quality recommendations for data creators

The recommendations listed below are designed to make the best quality collective data.

- Provide rich enough descriptive metadata for disambiguation purposes now and, if possible, a little more for future disambiguation.
- Provide as much non-sensitive data as possible for identities having a connection with your organization, including organization affiliations for personal identities.
- Provide unique and persistent local identifiers to provision for data synchronization, linking and diffusion of corrections.
- Reuse existing authority data (e.g., from NACO, VIAF, ISNI) whenever available and include external identifiers in your data along data being added locally.
- Do not make redundant data available; only share a local identifier when new data is created for that identity or your organization has a significant collection related to the identity
- Include relevant external identifiers as links (e.g., ORCID, MusicBrainz, Wikidata, VIAF, IMDb), when known.
- Capture disambiguating data in fields/elements that are machine-actionable; use controlled vocabularies consistently.

- Adhere to documented data models and application profiles
- Maintain provenance of the data, including source and revision history
- Identify and publish a measure of data confidence.
- If sensitive data is included, encode as such and only provide that data to trusted parties
- Use identifiers to disambiguate against similar names when creating directly within a shared authority file

For pragmatic reasons the above recommendations cannot be imposed systematically on existing data, but are strongly advised for newly created data.

The list of identities below, assembled manually from VIAF, ISNI, Google, MusicBrainz, IMDb, ResearchGate, ORCID and other internet sources, illustrates the importance of disambiguation for the relatively common English-language name Russell Thomas. The same name string is shared by three musicians, two film directors, and four authors who have written about education, retail, organic chemistry and environmental science. Many more distinct identities have the same inverse surname and forename Thomas Russell.

1. Thomas, Russell Film director http://www.imdb.com/name/nm1306805/
   Works: Coldplay: live 2003, Really bend it like Beckham (2004), Bill Bailey Tinselworm
2. Thomas, Russell Brown, 1900- writing on education
3. Thomas, Russell Tenor from Miami http://www.russell-thomas.com/bio.asp
4. Thomas, Russell – professor of Music and Director of Jazz education at Jackson State University http://www.jsums.edu/music/faculty/dr-russell-thomas/
5. Thomas, Russell J. 1966 writing on organic chemistry
6. Thomas, Russell N. 1973
7. Thomas, Russell Linwood, 1935 changed name to Al-Hajj Sayyd Abdul Al-Khabyyr – saxophonist
8. Thomas, Russell A. P. University of Strathclyde and Parsons Brinckerhoff – civil engineering and environmental science
9. Thomas, Russell A. Writing on retail
10. .....

When creating new authority records manually, data creators are encouraged to consult large data aggregations and use their identifiers in their local records to actively disambiguate their identity. Disambiguation is encouraged strongly, but it is difficult to impose and undesirable and unfeasible to police, especially for already established data. If active disambiguation is not feasible, efforts should be made to include detailed affiliation information to facilitate future disambiguation decisions, such as institutional or department affiliation.

The amount of data needed for disambiguation varies with the "commonness" of the name. Identifiers serve as a form of shorthand for the ensemble of metadata that is collected to describe and differentiate an identity. They are thus the key element in expressing links and in disambiguating an authority record.

5.3.1.4 Quality recommendations for data aggregators

- Use matching and merging algorithms that create duplicates rather than mixed identities
- Use provenance and data confidence statements provided by the data contributors in the matching and merging algorithms
- Provision for preferred clusters in data modeling to facilitate suspect duplicates to prevent the non-preferred authority data from attracting new data
- Provide unique and persistent cluster identifiers
- Restrict confidential information
- Indicate provenance of individual pieces of information
- Indicate data creation and maintenance dates, alongside data confidence when available
- Facilitate human-assisted or crowd-sourced workflows for enrichment, data correction and cluster correction
- Provide an automated mechanism for resolving data conflicts, e.g. encoding of pseudonyms and date anomalies
- Sample the database at intervals for quality. For example, the ISNI Quality Team samples every quarter to measure the level of duplication and mixed identities and to pinpoint areas of the database where manual intervention would be most profitable

5.3.1.5 Data quality summary

Quality issues that are unique to library authority data arise primarily because of the need to associate descriptions to real-world people and organizations and secondarily, to creative works, places, and topics or concepts. Since the challenges for data creators and data aggregators are largely distinct, the respective lists of recommendations shown in Section 5.3.1.3 for creators and 5.3.1.4 for aggregators have relatively little overlap. But creators and aggregators do interact; for example, a data aggregator may send a report to the data creator, which prompts the need to verify and correct suspected duplicates, mixed or ambiguous identities, and date errors.

5.3.2 Synchronization and diffusion of data and corrected data

Identity data can be highly dynamic, frequently changing as new data about the identity becomes available; once distributed, there needs to be a mechanism for updating all distributed copies. Keeping distributed data fully synchronized is impossible without online, real-time update as data is enriched and corrected. Yet universal real-time update is increasingly unrealistic as the number of files involved increases. Time delays can be significant. For example, a change made in the NACO file may take months before it reaches ISNI via VIAF. Where different versions are available, it is important to be able to recognize the current authoritative version via clues such as creation and revision dates, or provenance and maintenance responsibility. Use of persistent identifiers is also important for the quick and accurate location of a record to be updated.

Methods of achieving synchronization vary. VIAF currently accepts full replacements from its contributors and releases its full file for download at monthly intervals. VIAF also sends periodic

reports of possible error conditions. When full replacements are diffused, the responsibility is placed on recipients of the file to determine how to process them. For example, if deletions occur, the only way to detect them is to compare the new file with the previous. The file also needs to include deprecated identifiers so that merges can be detected. Other aggregations produce only updates or "diffs". For example, ISNI provides a regular notification service and an alignment comparison service on-demand. Once data is diffused, enrichment may occur in any source dataset. As a result, synchronization needs to be two-way – from the authoritative source to the aggregations and vice versa.

Aggregations also need procedures for synchronizing with other aggregations. Between VIAF and ISNI there are special procedures to ensure that merges, splits and corrections made by the ISNI Quality Team are also enacted in VIAF via the mechanism of "XA (extra authority)" records[18]. A researcher updating their ORCID record can consult ISNI and copy data, including the ISNI identifier, into ORCID at the same time as updating ISNI with the ORCID identifier; at present, the identifier updates in both directions are semi-automated, other data must be manually copied although this could be automated.

In a linked data world, misalignment among the key aggregators is seriously disruptive. Wikidata produces a misalignment report, but for ISNI this has not been very useful because the report largely consists of alerts to the assignment of different identifiers to different pseudonyms of the same identity.

5.3.2.1. Synchronization recommendations for data creators and data aggregators

- Data aggregators should provide at least one automated or semi-automated method for checking synchronization across systems
- Data aggregators should provide at least one automated or semi-automated method for notifying data contributors of data corrections to their own data and of cluster changes affecting their data
- Data creators should react to synchronization and reports on data corrections and cluster changes
- Data aggregators should provide information about cluster and identifier changes on a frequent basis
- Data aggregators should adopt necessary procedures to ensure alignment with other key aggregators
- Data aggregators should publish the acceptable formats and data models for ingest
- Data aggregators should document and make readily available formats and data models for published data

5.3.2.2. Summary

[18] For more discussion on XA records, see: MacEwan, A. (2016). *ISNI and VIAF: authority files and identity management*. Authority Data On the Web, Dublin, OH, 2016-08. https://www.oclc.org/content/dam/oclc/events/2016/IFLA2016/presentations/ISNI-and-VIAF-Authority-Files-and-Identity-Management.pdf

The publication of data on the web has made changes in the obligations of data creators and data aggregators, though the ones for the latter group are more significant. For data creators, there has been a shift from creating authority records that fit within an individual catalog to those that fit internationally, and with it a change from making a record to describing and differentiating an identity. For data aggregators, it is no longer enough to simply harvest and aggregate data from different sources, because identities must also be resolved. Aggregations not only provide online access, but distribute and carry the responsibility of distribution and synchronization of distributed copies.

5.4. Case studies - Aggregators

As previously discussed, aggregations play an essential role in data creation, data distribution and synchronization; without them, the tasks of differentiation and data correction are close to impossible. The current web environment has had a significant impact on data aggregators in the ways that data are collected, the amount that is collected, the number and diversity of contributors, and the diffusion of data.

Below are four case studies of aggregations, discussing issues of scale, quality and synchronization.

5.4.1. ISNI

5.4.1.1. Scale

Among other functions, ISNI acts as a cross-domain hub linking to VIAF, NACO, Wikidata, ORCID and rights management societies (e.g., Irish Copyright Licensing Agency, CEDRO (Centro Español de Derechos Reprográficos), and COPYRUS - Russian Rights holders' Society for Collective Management of Reprographic Reproduction Rights) among other sources; further, ISNI includes many individual libraries among its members and data contributors. The database is broad in scope with strong representation of identities associated with bibliographic works, research papers and theses, and musical compositions. Data can be contributed as files for batch loading or interactively, online via a web form or via a machine to machine API. ISNI also allows rights management and commercial sources to contribute data that is available for matching but not publicly available; this enables those sources to be a part of the system where otherwise privacy issues and business interests would keep them separate.

Unlike other aggregate authority files (e.g., NACO and VIAF), ISNI actively seeks input from the general public. Crowdsourcing is available via the "yellow box" in the online web interface and attracts enrichment and correction. This is an important element of ISNI, harnessing a wide-range of expert data and providing the world at-large the opportunity to participate.

The image below provides two examples of crowdsourcing input to the ISNI database, one requesting a merge and the other requesting splitting to differentiate two identities sharing the same name string. The ISNI Quality Team (QT) resolves these queries.

5.4.1.2. Quality

As a hub, the ISNI environment and system includes a comprehensive set of functions and processes facilitating data quality through swift, permanent correction of data errors.

The database is maintained online and curated by the ISNI Quality team at participating national libraries, led by the British Library and the Bibliothèque nationale de France. The team monitors and responds to all data provided by the public. Further, the team regularly samples batch contributed data and areas of the database to test quality. Reports derived from the aforementioned sampling serve to pinpoint enhancement needs within existing matching algorithms; these findings are shared with VIAF. Provenance of all data is recorded as is the level of confidence in the data. Different rules apply to data captured online versus data loaded in batch mode; in online mode, records that do not match may be assigned an ISNI if they are rich or if the name form is unique, rather than being typed "provisional". The provisional assignment aids in avoiding less complete and less authoritative authority data from creating duplicate ISNI assignments; records are not created in the case that insufficient data is provided. For more explicit information regarding ISNI's standards, see the Data Quality Policy statement[19].

The persistence of ISNI identifiers is built into the ISNI system. Where two records are merged, one identifier is deprecated, yet still resolves to the newly merged record. In the case that a

[19] ISNI (n.d.). *Data Quality Policy*. http://www.isni.org/content/data-quality-policy

record is split, a new identifier is created and the record of the previously assigned identifier will refer to the new identifier.

The aforementioned quality control algorithms are complemented by programs detecting errors and normalizing data, run at regular intervals. There is also a suite of special indexes designed for detective work, such as seeking unusual source combinations. ISNI achieves scale without compromising data quality by combining machine processing with manual input and control.

5.4.1.3. Sharing, Diffusion and synchronization

Data contributed to batch-create ISNI records are in either the ISNI XML schema or the ISNI tab delimited format. Special ingest is performed for VIAF in MARCXML and other sources such as MusicBrainz' schema[20]. The formats and ISNI's data element values are published on its web site [21]. There is a publicly available website and API delivering data in HTML, XML format and RDF as well as a downloadable file will be made available. Preparations are underway to publish the data element values in the form of a machine-readable ontology along with the RDF schema.

Once assigned, diffusion of ISNI identifiers and the publicly available metadata is encouraged. The ISNI request API enables contributing sources to retrieve and replace data in real-time, thus keeping their databases aligned. The source code of data contributors is used by the ISNI system to push notifications of any changes involving data correction such as merging or splitting. Contributors are also notified of new sources sharing the same data. There is a suite of comparison programs that enables an ISNI source to compare local identifier, ISNI and VIAF cluster identifier.

Although interactive update among aggregations is difficult to implement, Synchronization is best when updates are in real-time. Delays in updates reaching ISNI from secondary files has resulted in several libraries contributing directly to ISNI or becoming members so that critical changes can be streamlined. Between VIAF and ISNI there is a special procedure to accept work by the ISNI Quality Team and a regular exchange of potential matching improvements. VIAF ingests ISNI as a source with special treatment, such as ignoring data sourced from VIAF that may be missing an update; ISNI contributes all records having a VIAF or ISNI source code.

5.4.2. NACO

5.4.2.1. Scale

NACO is a Program for Cooperative Cataloging (PCC) program through which participants contribute authority records for personal, corporate, and jurisdictional names; uniform titles; and series headings to the LC/NACO Authority File (also called the NACO file, the National Authority File, or NAF). In Fiscal year 2017, there were over 900 individual NACO members primarily in English-speaking countries. NACO is intended as an English-language authority file, meaning that users have facility with the English language, although NACO records

[20] *MusicBrainz Schema*, accessed 2018.03.25. https://musicbrainz.org/doc/MusicBrainz_Database/Schema
[21] ISNI. *Resources*, accessed 2018.03.25. http://www.isni.org/content/resources-0

themselves can be contributed for entities in any language. Membership in NACO is open to individual institutions willing to support their staff through a process of training, review, and direct contributions of records to the LC/NACO Name Authority File.

There are 4 NACO nodes (also known as distribution partners)[22]: the Library of Congress, OCLC, the British Library and SkyRiver. Every day, the Library of Congress coordinates the updates from the other 3 nodes. The nodes themselves provide online maintenance interfaces, with most of NACO's 900 member contributors using the OCLC node via the Connexion client.

As of March 2018, there are 9 million NACO records, including approximately 6.2 million persons, 1.25 million organization records and 1.5 million name and title authority records.

5.4.2.2. Quality

The NACO database is maintained by the Library of Congress. NACO members may update, edit, or correct any NACO record via their affiliated NACO node. With 900 NACO members, there are a lot of trained eyes consulting and editing the database each day. This in itself contributes to the overall quality of the database. NACO responds to input from the general public (currently an average of 75 per month), via the Library of Congress online catalog. Further, NACO members may report errors to naco@loc.gov, a dedicated account for members. Because of the processing mechanisms for NACO records, only LC staff may delete NACO records. The PCC Secretariat at LC is the public-facing office for NACO. The PCC Secretariat consists of four full-time employees who process

- NACO maintenance requests (including delete requests from NACO members)
- Correction requests and new record requests from non-NACO members
- Machine-generated NACO error reports resulting from the FTP distribution process
- Answer general NACO questions for members

NACO records are identified by a Library of Congress Control Number (LCCN). When two NACO records are merged or when one NACO record is deleted in favour of another (or more) NACO records, deprecated LCCN's are maintained.

It is possible to include external identifiers (e.g.: VIAF, ORCID and ISNI) in NACO records; however, there is no systematic ingest of such external identifiers though an ISNI ingest is being considered (as of March 2018).

Currently there is no specific encoding in MARC 21 to indicate level of confidence in the data; MARC 21 is the underlying format used for NACO. Since it is desirable to be able to indicate high confidence (e.g., if the data were derived from personal correspondence), this is recognized as an area for consideration for future format enhancement.

A MARC organization code is included in the 040 field of each record to indicate the data provenance and the LCCN identifier indicates in which NACO node the record was first created.

[22] PCC (n.d.). *The NACO FTP process*, accessed 2018-03-28. https://www.loc.gov/aba/pcc/naco/nodes.html

Recent changes to the MARC 21 format to reflect RDA (Resource Description and Access) principles[23] have included new coded fields such as dates, associated places and languages. These new fields include data in a systematic way that was previously recorded in free form in a note field. Although some of the new fields in name/title records have been machine-generated from other data in the record, converting legacy data is a challenge not yet fully addressed.

Unlike the aforementioned ISNI use case where private data is retained for matching purposes, all NACO data are available publicly. Strict regulations exist about the data to include in a record so that privacy can be respected. For example, address information (not including email addresses) is never recorded in NACO records for living persons.

The NACO database includes undifferentiated name identities; in other words, not all NACO records reference an individual.

> Sometimes two people with the same name cannot be distinguished. Their birth and death dates cannot be determined from reference sources and no other information can be found to break the conflict. In these cases, catalogers will generally create what is called an "undifferentiated personal name" authority record. The same record (and the same heading) will be used to represent more than one person in the catalog.[24]

Training for NACO catalogers is intensive with an emphasis on comprehensive searching before a new record is made. This minimizes duplication but does not eliminate it. Duplicates may occur when a record for the same identity is made on the same day in two different NACO nodes. The daily consolidation of the input from each node checks LCCN and date/time stamp; however, there is no duplicate detection.

Daily reports signal invalid characters, incorrectly coded scripts, incorrectly coded MARC 21 tags and subfields, and multiple record updates on the same day. In order to ensure version control of records loaded into the LC master file, a program check is set on the MARC 21 005 field (Date and time of latest transaction). The 005 field contains the date and time (down to a tenth of a second) of the latest transaction, which ensures that the change coming into the database has used the latest version of the record in the master copy of the name authority file at LC. This also assures that an earlier version of a record will not replace a later version of the same record.

5.4.2.3. Sharing, Diffusion and synchronization

NACO nodes must hold a copy of the 9 million-record LC/NACO Name Authority File (NAF), and they must keep it current by loading daily NACO distribution files from the Library of

[23] Library of Congress, MARC Standards (2009). *MARC 21 Format 2009 Changes to Accommodate RDA (Draft)*. https://www.loc.gov/marc/formatchanges-RDA.html
[24] Yale University Library (2015). *Undifferentiated Name Authority Records*. Cataloging@Yale. https://web.library.yale.edu/cataloging/authorities/voyager/tips-undifferentiated-name

Congress Cataloging Distribution Service (CDS). Thus, the copies held by the nodes are mostly up to date to the last 24 hours, maximum to the last 48 hours.

Nodes must also retrieve and process daily response files that contain error messages that pertain to records that were contributed to LC the previous day. Distribution recipients generate their own LCCNs, each with a distinguishing prefix (i.e., n = LC; nb = BL; no = OCLC; nr = RLG (no longer active, although nr records still reside in the NAF); ns = SkyRiver).

Each day LC retrieves a single contribution file from each distribution recipient and processes the files immediately. These files contain both newly created and changed name authority records.

The NACO file is available through the Library of Congress Linked Data Service http://id.loc.gov/ in a large number of serialisations: RDF/XML (MADS and SKOS), N-Triples (MADS and SKOS), JSON (MADS/RDF and SKOS/RDF), MADS – RDF/XML, MADS – N-Triples, MADS/RDF – JSON, SKOS – RDF/XML, SKOS- N-Triples and SKOS – JSON. The NACO file is an integral part of the BIBFRAME pilot.[25]

OCLC ingests the NACO file into VIAF from which is it widely distributed in MARC 21, XML, RDF and JSON. ISNI ingests VIAF and thus includes NACO records, with the exception of name/title authority records; name/title records are out of scope for ISNI, which focuses on person and organization identities.

5.4.3. VIAF

5.4.3.1. Scale

The goal of VIAF is to consolidate authority files from the world's national libraries and leading research and cultural institutions, creating a freely available reference tool. The VIAF website receives more than 1 billion hits per year and its data is downloaded more than 80,000 times per year. VIAF plays an important role in the emerging semantic web, being a highly used and cited resource.

VIAF does not actively seek public contribution but does receive a small volume via the "Send us a comment" function on their home screen.

5.4.3.2. Quality

The sources of VIAF data are responsible for data creation and maintenance. VIAF harvests the data and forms clusters, preferencing duplicate clusters rather than creating mixed identities when a match is not confident. VIAF creates separate clusters where there are data conflicts,

[25] Library of Congress, Acquisitions & Bibliographic Access Directorate (2016). *BIBFRAME Pilot (Phase One—Sept. 8, 2015 – March 31, 2016): Report and Assessment*. https://www.loc.gov/bibframe/docs/pdf/bibframe-pilot-phase1-analysis.pdf

such as date conflicts. All data are exposed and there is no hierarchy of clusters; however, undifferentiated data derived from NACO is flagged and left un-clustered.

Undifferentiated records from the Deutsche Nationalbibliothek are included if a match is identified and the data are not sparse. Other undifferentiated records are not included in the VIAF database. Sparse data are also flagged and about half of the records are un-clustered. Encoding differences are tolerated such as the encoding of pseudonyms. Hickey and Toves (2014) provide a detailed discussion of managing ambiguity in VIAF[26]. VIAF does not normalize source data; however, special records can be made manually to overcome ambiguities and facilitate clustering. VIAF can also force clustering or prevent incorrect clustering using XA (extra authority) records. Manual merges and splits conducted by the ISNI Quality Team are flagged such that they have the status of XA records during VIAF clustering.

Special procedures exist for maintaining the persistence of VIAF cluster identifiers; although in the case of multiple clusters for the same identity (for the reasons explained above), source data may move from one cluster to another. A history of each cluster is maintained and can be viewed online.

Wikidata is imported as a VIAF source with the exception that only matching data is retained. A close cooperation with Wikidata and Wikipedia exists; as of March 2018, OCLC Research is framing a project for the automated detection of anomalies and their subsequent correction.

5.4.3.3. Sharing, Diffusion and synchronization

VIAF is available as an online resource and as a downloadable file in RDF, XML, JSON and MARC-21 (variation) formats. As of March 2018, documentation of these formats is in-process. MARC 21 was adapted to suit clusters of separate records, particularly allowing repetition of the established name. VIAF sources can ingest VIAF IDs and enriching data into their databases using the monthly full VIAF file. This file is also used to monitor synchronization of local data with VIAF. For example, ISNI compares new VIAF downloads with the previous download; from the comparison, ISNI produces sub-files of additions, source changes, content changes and deletions.

Occasionally, VIAF produces reports for data contributors signalling possible errors, including potential duplicates in the local file and local records that are in singleton VIAF clusters but may be duplicates.

As mentioned above, special procedures have been devised to align ISNI and VIAF.

5.4.4. SNAC

4.4.1. Scale

[26] Hickey, T. and Toves, J. (2014). *Managing Ambiguity In VIAF*. D-Lib Magazine 20 (7/8). http://www.dlib.org/dlib/july14/hickey/07hickey.html

Initially established as a research and demonstration project, the Social Networks and Archival Context (SNAC) project is currently transitioning from project to program with the establishment of the SNAC Cooperative. During the creation of the database, records were generated based on information about archival creators extracted from archival descriptive records. Through this process, approximately 3.7 million entity records were generated, having been harvested from 150,000 contributed EAD-encoded finding aids, 2.2 million MARCXML records for archival collections in OCLC WorldCat, and 375,000 contributed archival authority records.

At the time of its establishment, the SNAC Cooperative consisted of 17 Cooperative members, including research libraries, the Library of Congress and the National Archives and Records Administration. Ten more contributors were added for the second phase of the pilot, increasing the diversity of the participants.

5.4.4.2. Quality

The quality of the generated authority records available in SNAC is dependent on the source records contributed to the project. During this process, entity names were extracted from existing records, and matched against each other and VIAF. Duplicate records were then identified and merged to produce a single EAC-CPF record for each entity.

The semantics for batch processing the harvested data was primitive, so only a certain level of precision was possible. Recorded relationships were especially basic, with most defaulting to an "associatedWith" relationship designator. During 2016-2017 the SNAC Cooperative established an Editorial Policy and Standards working group to bring greater consistency to the contributed data, and to develop guidelines for its ongoing maintenance.

As of early 2018, the project is working to develop an online Editing User Interface (EUI) for creating and updating entity records. Contributors will also be able to submit changes, as well as new records, through the API. Work on merging and splitting identity clusters is ongoing.

Some ethical and privacy issues have emerged with the development and publication of the SNAC dataset. These include concerns about identity theft, as well as social issues such as gender identity. The general policy thus far has been to defer to the wishes of the individuals represented in the authority records, insofar as it does not lead to mischaracterizations. The site provides a user feedback link on all entity records, allowing users to request changes as needed.

5.4.4.3. Sharing, Diffusion, and Synchronization

The SNAC project team is currently working on revising the public interface, which provides researchers with direct access to the dataset. Records are downloadable in either the EAC-CPF or JSON formats. It is expected that RDF downloads will soon be available for a subset of the data using the JSON-LD format. There are also plans to develop a SPARQL endpoint service, and to allow batch downloads of the entire dataset.

The dataset includes ARK identifiers to each authority record; ARK identifiers are a type of persistent identifier[27]. Policies may need to be developed to allow persistent identification in the case of record splits.

Within the dataset, each authority currently includes links out to VIAF and to DBpedia. However, there are some concerns about maintaining these links, and they are working to develop technical solutions for identity verification. The project would also like to develop a direct relationship with Wikipedia to allow the automated insertion of links to SNAC records into Wikipedia pages.

SNAC is also working to develop tools for automating the ingest and synchronization of records contributed by Cooperative participants. This will include an identity disambiguation service, which would allow contributors to query the database to verify whether a record already exists for an entity before a new record is created.

6. Workflows

6.1. Introduction

This section provides a generalized overview of six workflows relevant to both data creators and data aggregators. The intention of section 6.1 is to demonstrate generalized workflows unrelated to any particular service. Although the following workflows are intended to be general, ISNI is used throughout the aggregator workflows to demonstrate how these are handled in that environment.

Following discussion of these six workflows, five cases studies are provided.

Workflows related to a data creator perspective:

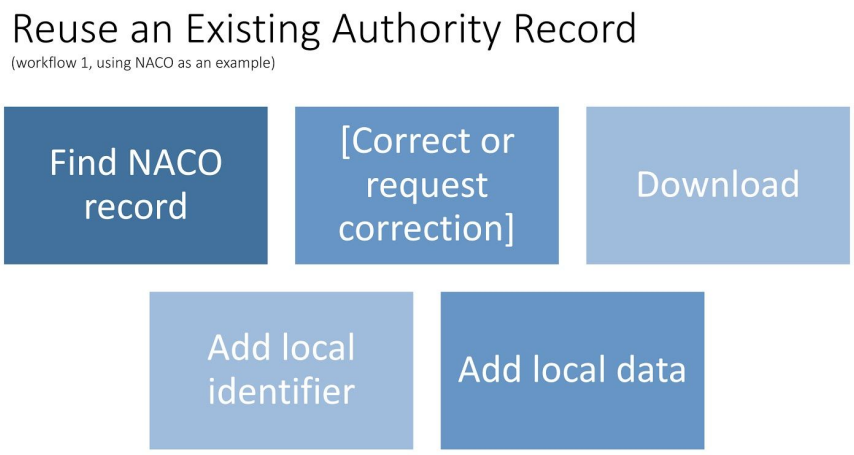
- For ongoing work, reuse an existing authority record
- For ongoing work, create a new authority record

Workflows related to a data aggregator perspective:

- Contribute authorities to an aggregated source
- Diffuse corrections from a collective file
- Synchronization with a collective file
- Crowdsourcing

[27] For more about ARK identifiers, see: Kunze, J., & Rodgers, R. (2008). *The ARK Identifier Scheme*. UC Office of the President: California Digital Library. Retrieved from https://escholarship.org/uc/item/9p9863nc

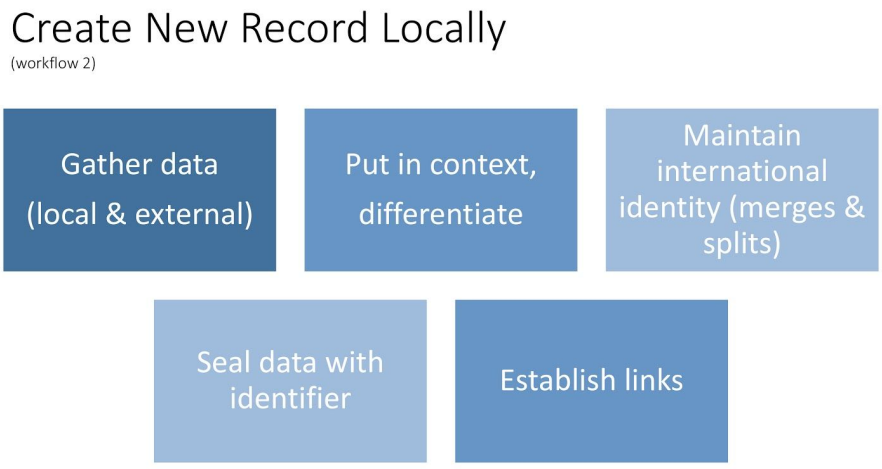
6.1.1. Reuse an existing authority record



Sources such as the NACO file, VIAF, ISNI and WorldCat identities are searched to see if an authority record already exists. The sources may be consulted via the service's websites or via the data creator's local system using an API or other data exchange mechanism. If an appropriate record exists and considered to be good data, the record may be downloaded into a local system and a local source and identifier may be added to the central record to enable future data exchange. Enhancements may include additional metadata or links to related persons or organizations; links may be among database records or external URL links.

If the record is deemed not to be good then, depending on the system, the library may make corrections, merges and splits - or request such changes be made within the source data.

6.1.2. Create a new authority record



When an existing identity does not exist in a local authority database or in an aggregated file, a data creator must mint a new identity. The following workflow is irregardless of whether a data creator is working in a local versus shared environment. Data for authority records may be gathered directly from the person or organization concerned, from local sources (e.g., administrative records or repository data) or from external sources (e.g., encyclopaedias, biographies, professional associations, research sharing, social networks and websites). Contextualizing data and disambiguating the identity is as crucial as gathering information about the individual.

The source of the new authority record should assume ongoing responsibility for the record. This involves period checks to ensure that the record is still correct and correctly disambiguated, if enhanced in an aggregated file.

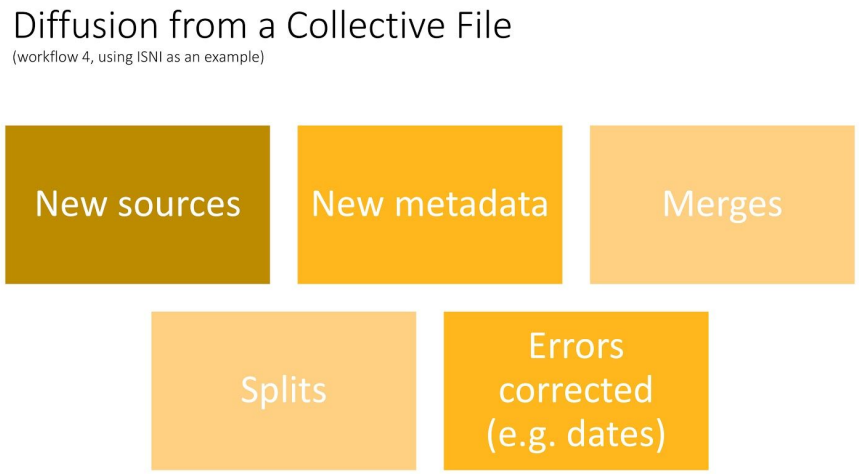
6.1.3. Contribute Authorities to an Aggregated Source

When contributing a batch of local authority records to an aggregated authority file (e.g., ISNI), reconciliation efforts must occur between these two datasets (see section 5.3.1.1 and section 7 for more details). Following batch loading to aggregated files, reports may be produced indicating records for which manual review is desirable. To ease reconciliation and reduce potential data conflicts, data creators should provide sufficient data (see section 5 for more details).

In the case of ISNI, reports identify where a match confidence score between two identities was less than that required for automatic matching. Online queries are run to check matches with unlikely sources to avert mixed identities erroneously created by the algorithms. An online query can also be used to identify the records that have been assigned because the name string is unique to the file. These records undergo review to avoid assignment of names with spelling errors.

Data creators must assume responsibility moving forward for the data contributed to aggregated sources; this incurs checking future reports (e.g. of end user contribution) and spot checking for enhancements.

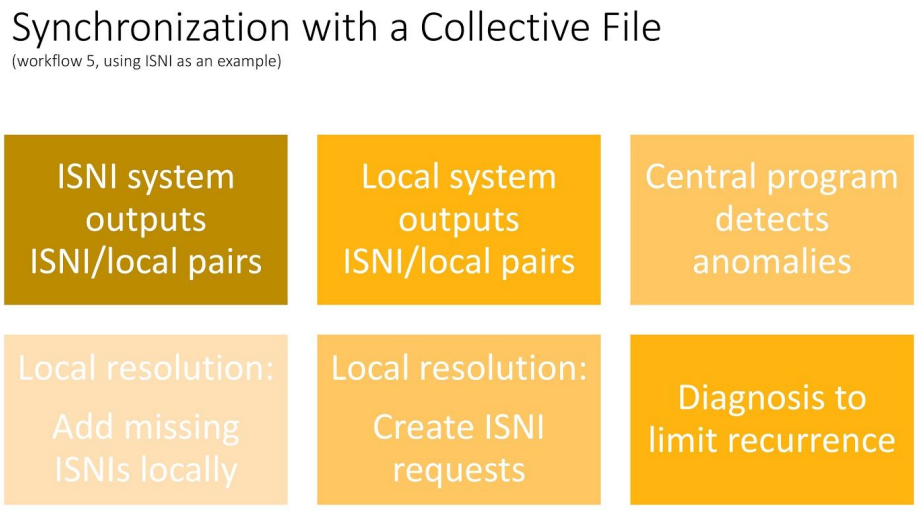
6.1.4. Diffusion from a collective file



At times, data shared contributed to an aggregated source may incur enhancements that are relevant to the local authority file in which the data originated; this can include straightforward enhancements (i.e., additional data) as well as complex logic (e.g., identity merges or splitting multiple identities from a single identity). In this case, data creators may wish to take action on their local data.

In the case of ISNI, regular reports are sent to data contributors indicating various database activity on records in which your source occurs. This includes notification of new sources and / So new metadata added to a record, merged records (indicated the retained and deprecated identifier), split records and data corrections, such as dates.

6.1.5. Synchronization with a collective file



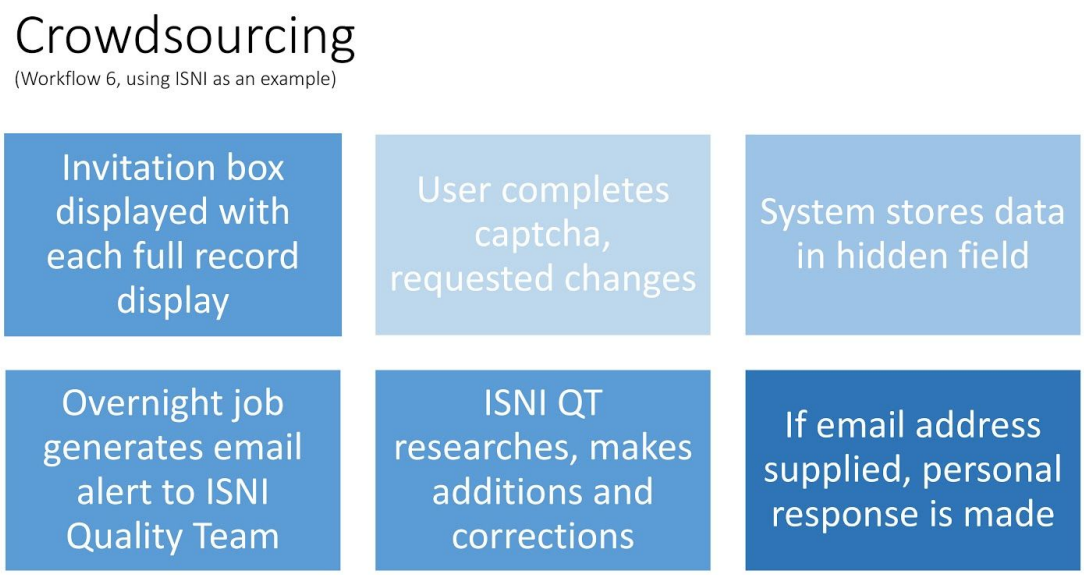
As discussed in section 5.3.2, synchronization is a fundamental concern when local authorities are contributed to aggregated services.

Synchronization tests exist in the ISNI system whereby ISNI produces a list of ISNI identifiers and local identifiers that is compared with a similar list created from a contributing file. Lists of anomalies are then made where missing and misaligned data are detected. Corrections are made along with an analysis to find the origins of the error in order to find ways to prevent recurrence.

Wikidata also produces lists of anomalies but in the case of the ISNI report, most are due to ISNI assigning different identifiers to a real name and a pseudonym.

VIAF and ISNI have special measures for maintaining synchronization. VIAF includes special records called XA (discussed in section 5.3.2) that have special weight and meaning for the merging algorithms. These are used to favour ISNI merges and splits.

6.1.6. Crowdsourcing



To different extents, ISNI, VIAF, SNAC and NACO all have functionality built to facilitate crowdsourced data quality assessment and data edits.

ISNI has a special workflow for contributions made by the general public. The data collected from the web form are entered into a hidden note field to prevent incorrect or malicious data becoming visible. A nightly process searches these new fields and generates email reports for the Quality Team. Necessary corrections and enhancements are made real time and an email response is also sent to the contributor if an address had been supplied. More libraries are joining

as associate members of the Quality Team ensuring that the team will be able to cope with future increases in the number of contributions.

6.1.7 Issues that arise in aforementioned workflows

How much metadata is needed for differentiating and preventing future errors?

Traditionally authority records were made with recommended levels of completeness in the hope that the more complete the record, the more it would be future proof. The amount of data actually required for disambiguation varies by the rareness of the name. Even for rare names, it is desirable to record such detail as titles of several works and co-authors to make it easier to determine if a new work is by the same identity. Sometimes explicit "is not" statements separating persons of the same name are more useful than any other metadata.

Which institution is responsible for an authority record when the person is associated with multiple institutions?

This can be argued. The institution with which the person is most recently associated is more likely to identify and add new works but the institution that first contributed could be logically responsible.

How to reconcile different data styles, e.g. recording pseudonyms

One example is the treatment of pseudonyms. Wikipedia collects all pseudonyms together with a person's real name. Libraries vary; some treat pseudonyms as separate identities and some as name variants. Rights Management Societies almost always treat pseudonyms as separate identities. In ISNI, programs have been written to identify pseudonyms flagged as name variants to convert them consistently to related names. These programs only have a limited effect as not all pseudonyms are flagged as such and those that are, are flagged in multiple languages and abbreviations are common.

6.2. Case Study: British Library

Contributed by Andrew MacEwan (British Library)

The NACO file is loaded into the British Library's internal Aleph system, and serves as the major part of the authority file for the British Library (BL) catalog. A new authority record is made where there is a unique name; where it probable that the identity's publication is their first; or where subjectively, none of the existing authority records seem to represent the same identity (based often on subject of publication and date of publication). Where there are possible authorities, research is made via the web, email and telephone to establish whether an identity is new or already existing, enhancing both existing and new records in the process. When a new authority record is created, the system automatically populates title, name and other fields. Further data is collected manually from multiple sources as deemed necessary including, but not limited to, biographical information supplied with the work, VIAF, ISNI, ORCID, encyclopaedias, web pages found via Google searches, published indexes, emails and phone conversations with publishers, authors and other creators. Data are exported to NACO. The

British Library is one of the major contributors to NACO. Via NACO, the data are contributed to VIAF and thence to ISNI. BL derives NACO, ISNI and VIAF IDs via the NACO load. The BL makes its authority data available via Z39.50 searching and as linked data under the CC0 license using bnb.data.bl.uk. Two different interfaces are provided: a SPARQL editor, and /sparql a service endpoint for remote queries with numerous output options: Linked Open BNB, Basic RDF/XML, Researcher format (CSV), PDF and MARC 21.

The BL authority file also includes authority data records that are not exported to NACO. Authority records are made for Electronic Theses (EThOS). The criteria for making a new record are the same as for NACO work. EThOS data is loaded to ISNI, but is not included in NACO. ISNI provides regular load reports that are used for populating the BL authority file with ISNIs.

The British Library and the Bibliothèque nationale de France lead the ISNI Quality Team. The work involves sampling for duplicates or errors and responding to end user input coming from the open ISNI database enquiry page. The samples are scheduled each quarter and additional occasional samples are made. The regular samples use queries largely based on source contributor but occasional samples can use any of more than 20 indexes available for search and browse. End user input flows regularly at a current rate of more than 150 a month. Enhancements and corrections are made directly in the ISNI database using online tools. Corrections made in the ISNI database are automatically diffused to all concerned sources, however the BL team systematically repeats the corrections in the BL authority file and NACO where appropriate (as does the BnF systematically repeat corrections in its national file). ISNI provides specific reports on request that include NACO, VIAF and ISNI IDs and the combinations are compared with previous reports.

The Andrew W. Mellon Foundation is funding the development of a portal at the British Library that is being designed to attract crowd input and will at the same time coordinate the three separate workflows and databases, described above, as it is phased into operation. The ISNI database will be the point of data creation and maintenance, flowing to NACO. The target market will be the UK public but availability of the portal will not be restricted. End users will be able to directly enhance ISNI data and make ISNI requests via specific forms, giving an experience superior to the current ISNI interface that allows a hidden note to be conveyed to the ISNI Quality Team.

6.3. Case Study: Harvard Library

Contributed by Michelle Durocher and Chew Chiat Naun (Harvard University)

Name authority control for the Harvard University Catalog has historically been centered on NACO. The Library uses the MARS automated authority control service from Backstage Library Works to validate headings in new cataloging records against the NACO authority file and incrementally to update existing records. A team of cataloging staff works through a monthly report from Backstage to update headings and resolve problems. Authority records that match headings in the Library's Aleph catalog are stored in the Aleph authority database. Where new NACO authorities are needed, they are created and edited in Connexion, then exported to Aleph.

This workflow has been augmented by the creation of local name authorities which remain in the Aleph system and are not distributed outside Harvard. The practice of creating these authorities only in the local system has been driven by past system limitations and by barriers to certifying sufficient numbers of staff as NACO contributors.

During 2018 Harvard will migrate to the Alma system, which has an architecture that emphasizes the use of shared authority files. The Library is analyzing match reports to determine which of our existing local authorities need to be migrated, and is developing workflows to minimize or eliminate future creation of local authorities. In the short term these efforts will focus on increasing our capacity to contribute to NACO. In the longer term, the Library will be investigating identifier-based solutions that we hope will be more scalable and interoperable (see below).

Alongside the main Harvard catalog, there are other pools of authority data that are maintained in separate systems. To date only limited efforts have been made to interlink them. These include archival records maintained in ArchiveSpace, visual materials maintained in JSTOR Forum and researcher data maintained in the university's DSpace registry. The institution registry interfaces with the University administration system using LDAP.

Looking to the future, Harvard is working with ISNI and the PCC. The university is following the NACO lite[28] proposal that aims to ease the burden of creating new records, e.g. by eliminating the necessity to create a unique name heading string, and concentrating instead on encoding disambiguating information. A small pilot with the design school worked with ISNI and NACO, looking at ways to streamline the workflow and at data inconsistencies and systematic ways to overcome them. Loading to ISNI from the Institution Repository (IR) as was done at La Trobe University may be a way of uniting the IR data with the catalog data and other ISNI sources[29].

6.4. Case Study: Opaque Namespace

Contributed by Sarah Seymore (University of Oregon)

OpaqueNamespace (http://opaquenamespace.org/) is a local vocabulary manager with a Github backend where Oregon Digital local authorities are both stored and created. The records are stored as N-Triples and downloadable as JSON-LD or N-Triples. The workflow shared here provides an example of local authority work in a digital library repository.

Oregon Digital reuses existing authorities such as the Library of Congress Name Authority File (LCNAF) and Getty Union List of Artists' Names (Getty ULAN). The current workflow gathers data about names as items are being described and metadata is being applied. For a previous

[28] Durocher, M. and MacEwan, A. (2015). *NACO Lite? -- re-imagining Harvard's local name authority workflow as an identity management workflow using ISNI*. PCC Policy Committee Meeting, 2015-11-05. https://www.loc.gov/aba/pcc/documents/ISNI-NACOLite-PCC-PoCo-Nov2015.pptx

[29] ISNI (2015). *Member Story: La Trobe University*. http://www.isni.org/content/member-story-la-trobe-university

migration project, a programmer ran lists of names against LCNAF and Getty ULAN for reconciliation, which was marginally useful and still required individual URI confirmation. New authorities are created as a requirement for descriptive metadata in Oregon Digital. The metadata fields for creators and subjects (among others) in Oregon Digital require URIs. When no URI is found in an external vocabulary, it is created locally. Reconciliation is done at time of entry and there is no external reconciliation after an authority URI has been created.

Oregon Digital occasionally contributes to Getty ULAN via the webform application, and are hoping to engage further in making the process more regular and streamlined in the upcoming months, as there is a backlog of artist names to contribute.

The steps in creating a new authority in OpaqueNamespace are as follows:

- A new authority URI is created after other external linked data authorities have been searched and no resolvable URI has been found. In http://opaquenamespace.org/, an editor, reviewer, or administrator can create a new URI but must first determine the corresponding vocabulary for that URI: http://opaquenamespace.org/vocabularies. If a new vocabulary needs to be made, it is reviewed and approved first by University of Oregon (UO) and Oregon State University (OSU) for inclusion. Once the vocabulary is determined, the user enters the information for the URI in a webform.
- The fields are:
    - ID: The URI without spaces or punctuation; example: http://opaquenamespace.org/ns/people/SowardsSusan
    - Term Type: Concept, Corporate Name, Geographic, Personal Name, Title, or Topic
    - Label: The human readable label in "last name, first name" format that is displayed in Oregon Digital to the user
    - Comment: At UO, comments are added to provide as much contextual information about the person/concept/organization, etc. as available. This comment includes the collection for which the URI was created. Example comment: "Associated Students of the University of Oregon Staff active ca. 1920-1989 pictured in the UA Ref 3, University of Oregon Libraries, Special Collections and Archives photographs, 1890s-2010s collection."
        - The ID, Term Type and Label fields are required. Comment is not required by the system, but strongly encouraged in the workflow
    - Additional fields:
        - Alternative name
        - Date (birth-death)
        - See also
        - Is defined by
        - Same as

- Once this information is entered and submitted, the record enters a review queue and another person with an admin or reviewer role can approve the URI. It is then ready for use

There is a review queue process for creating and approving new URIs. Data for opaquenamespace.org is kept on Github. Broad statistics and branch information can be accessed (new/edited records create new branches that are merged when reviewed). URIs are approved by one other person at UO or OSU with an administrator or reviewer role. Editors can contribute URIs but those records have to be approved by an administrator or reviewer. If an external authority, such as LCNAF, publishes a term already in http://opaquenamespace.org/, the OpaqueNamespace record can be depreciated, with the new URI recorded in the record.

Continued development on http://opaquenamespace.org/ happens once or twice a year, as developer time becomes available. Although there are outstanding issues with functionality and ease-of-use, OpaqueNamespace reached Minimum Viable Product status in late-2017. Maintenance is continuous, with lead developers reviewing GitHub issues and making small changes as needed. Planned improvements for the future include the addition of relationship fields, automatic generation of IDs, refinement of the record change history, and streamlining bulk record ingest.

The Github repository for OpaqueNamespace is available at:
https://github.com/OregonDigital/ControlledVocabularyManager

6.5. Case Study: University of North Texas

Contributed by Hannah Tarver (University of North Texas)

University of North Texas (UNT) links to existing sources for authorities when available (e.g., LOC, Twitter, ORCID, Scopus, ISNI, etc.); however, information is not imported from these external sources into local systems. Most authority records are made locally for persons named on specific items that are added to UNT's Scholarly Works repository. Basic information (e.g., name, affiliation, and research areas) are taken from the item in-hand. Other data is added from sources including information from UNT's faculty profiles or other authorities.

New authorities are created due to a need to manage names that already had established authorities as well as names that were not yet controlled. UNT has particular concern for the creation of identities related to faculty members and others entities associated with UNT. These local names fall within UNT's purview and are unlikely to be controlled by other agencies. UNT decided the best way to handle this was by creating a local authority that assigns a new, local identifier for each name whether or not it has other authorities/identifiers; this also allowed UNT to directly connect our authority records to the metadata editing interface. UNT Digital Collections does not contribute to a collective file, however NACO work is done by other departments in the libraries. There is no set schedule for quality assurance or quality control testing; however the list of names is reviewed periodically, with most issues reported by metadata editors.

Authority records are stored in a database on a local server managed using an administrative interface connected to the UNT Name App (http://digital2.library.unt.edu/name/); records are publicly available in MADS/XML and JSON formats. Administrators can edit name authority

records any time. As of March 2018, the system cannot store the links in the records; thus, names in the records are simple strings and have to be changed manually on an individual basis to correct mistakes.

Steps in creating a new authority are as follows:

- A staff member with administrative access creates a new entry in the name authority database, containing any information known about the entity, and saves the record. At that point, it is immediately public and connected to the editing interface.
- At minimum, a new authority record includes :
    - Locally-authorized form of the name; generally, this is the most complete form of the name available, if the name does not already have a controlled version in another authority
    - Type (person/organization)
    - Affiliation for persons associated with materials in the repository.
- Additional information may be added at that time if it is available, such as date(s), alternate forms of names, links to other identities/authorities, and additional biographical information (e.g., name changes or research interests). When adding a large number of new authority records, the UNT Repository Librarian tracks the names added so that they can add information later.
- For organizations, UNT staff generally check for readily-available information in VIAF, Wikipedia, institutional websites, etc. Personal information may also come from authorities or from faculty profile pages.

6.6. Case Study: Western Name Authority File

Contributed by Anna Neatrour and Jeremy Myntti (University of Utah)

The Western Name Authority File (WNAF) is a regional database of personal names and corporate bodies currently being investigated with funding from the Institute for Museum and Library Services (LG-72-16-0002)[30]. This pilot project and future development will be hosted at the University of Utah. While the final format and tools used to host the database have yet to be determined, early stage workflows for name gathering, reconciliation, and deduplication are shared below.

Names are sourced from partner libraries[31], who submit data in a variety of formats. Names data has been submitted by full collection exports into spreadsheets, lists of names, and as JSON data. Data submitted is then transformed into csv format, merged, deduplicated, and reconciled against the LC Name Authority File. In the future we plan on exploring reconciliation against SNAC as well.

Names will be created primarily based on the representation of those names in regional digital collections. Data will be managed locally with open source software. During the pilot phase of

[30] Western Name Authority File (WNAF) grant application: https://www.imls.gov/grants/awarded/lg-72-16-0002-16
[31] Western Name Authority File Project (2018). *Project Participants*. https://sites.google.com/site/westernnameauthorityfile/project-partners

this project, we are making the data available using OmekaS. We anticipate a workflow where data is initially batch loaded into a name management system, with access to regional editors to that system established so they can also upload subsequent names directly.

Fields being collected for all names in the WNAF include:

- Preferred form of name
- Alternate form(s) of name (if available)
- Local authority source
- Institution holdings
- Relationship information
- Associated dates
- Geographic area
- Local identifier
- LCNAF URI (if available)

We plan on developing a regular maintenance and quality control plan once the database is online. In addition, we plan on contributing to NACO as names of particular interest are identified and researched. Workflows for adding new names to WNAF, reconciling new data against WNAF and other authority files, and making updates or corrections to the data are currently being investigated since we are in the pilot phase of the project. Project updates can be found at: https://sites.google.com/site/westernnameauthorityfile/

7. Reconciliation as a Service

7.1. Introduction

Discovering authority records for entities that are not closely controlled in local systems and do not meet the requirements of existing shared authority systems is an ongoing problem that limits the ability for dispersed items to be pulled together under one heading. A cultural heritage organization may have ancillary holdings of works by a particular individual as part of a larger collection (such as a letter received from an author that would benefit from a link to that author's finding aid of other artifacts). A library might seek to retrospectively add more prominently used headings to a collection of undergraduate theses projects to make them more discoverable. In both cases, headings may exist in the automation systems of other institutions (in the first case an ArchivesSpace installation that describes the recipient's letters and in the second case a DSpace institutional repository where that student-now-researcher has published new works), but those headings cannot be found.

During the in-person meeting at Cornell University on October 24-25, 2016, one of the concepts raised during the final panel was Reconciliation-as-a-Service (RaaS)– a stack of software and data that, when deployed, would:

- Harvest name authorities from local sources (e.g. VIVO installations, and institutional repositories) and aggregator sources (e.g. ISNI/VIAF/Getty)
- Use data from the harvested name authorities to cluster likely matches
- Provide a programmatic API for a user interface component to search/filter on the harvested data
- Provide a programmatic API for a user interface component to submit authority data to a few selected aggregator sources

As envisioned during the the in-person meetings at Cornell and the follow-on meeting at the Library of Congress, such a service would enable an institution to gather name authority data from an ad hoc list of published sources for its own reconciliation needs. As a locally hosted service, that institution could choose preprocessing steps and clustering parameters that result in the highest possible precision and recall for its particular needs.

7.2. RaaS: What It Is and What It Does

In order to gain a clearer idea of the envisaged Reconciliation Service, and to ensure that all participants in SLNA-NF shared the same conceptualization of its purpose and functionality, the authors of this document undertook two exercises frequently employed in software design: the development of personas, scenarios and of user stories.

Personas are idealized or imagined end-users, described in detail in order to assist in ensuring that the software product is designed in a user-centred way. In the context of SLNA-NF and its diverse participants, the use of personas also helped to ensure that a wide range of possible stakeholders was represented in the specification and design process.

Scenarios describe situations that end-users hope to accomplish; these detail the motivations behind end-user actions.

User Stories capture the needs of the end-user, employing words that are common in the business practices of the end-user. The user story describes the type of user, what they want and why, and often uses this template: "As a <role>, I want <feature> so that <reason>."

7.3. Personas

7.3.1. Becka is a skilled data-entry technician for a commercial company specialising in archive digitisation. As part of her work, she generates authority lists of all the people and places mentioned in the contents of an archive – and would like to align these lists both with each other, and with external authority files. At a technical level, she is extremely skilled in, and happy to learn, all the nooks and crannies of the software she's given to perform her role, such as the company's bespoke CMS. But she is not a developer, and requires a friendly UI and functionality if she is to do her job effectively. (Related to scenario 7.3.1)

7.3.2. Christina is a strong all-purpose developer. She's been through coding boot camp, worked in a variety of environments from QA to full-stack, and uses a lot of different tools and languages to get the job done. Right now she's working on maintaining cataloguing systems for a big college in Texas. In some ways she finds the work easy – the datastore is not particularly large by commercial standards, and load isn't heavy. But she doesn't really understand the domain, and is often a little puzzled as to *why* the campus librarians want to do what they do. (Related to scenario 7.3.2)

7.3.3. Darren is a metadata specialist at a large state university. He spends a lot of his time on what he sees as routine cataloguing tasks. But the university also has some archives and special collections he's got plans for – he'd like to finish the various digitisation projects that have been started over the years, and in particular he'd like to release some of the smaller archives as Linked Open Data, and in accordance with LOD best practice. But time and resources are short. There's no way he's going to be able to roll his own solution for this, particularly given his scrappy coding skills: he's done plenty of work with XML, and a little scripting, but it's the kind of thing he needs to relearn every time he tries it. (Related to scenario 7.3.3)

7.3.4. Joris is the digital curator of a large, publicly-funded gallery and museum in the Netherlands. He is a strong advocate for Linked Open Data as a means of publishing his institution's data and linking it with that of other museums. But his development resources are limited – and his stakeholders aren't eager to invest time and money in something they're not sure they entirely understand, that isn't easily demonstrated, and that is difficult to budget for with confidence. (Related to scenario 7.3.4)

7.3.5. Nathalie is a collections editor for a commercial publisher and aggregator: she finds materials, often archives, digitises them (if they're not born digital), organises them into a coherent whole, ensures their quality, and is generally responsible for curating and managing them. She's interested in LOD as a means of linking up her collections, organising them, and ensuring their discoverability. But her concern is very high-level: she has little technical background, and is often frustrated at the opacity and delays interactions with developers frequently seem to bring. (Related to scenario 7.3.5)

7.3.6. Nuno is a digital-libraries developer and repository manager for an elite university in Spain. He has extensive experience developing for university and research libraries, and a senior, well-funded position managing three other developers and librarians. His chief frustration is that his team spends so much time on routine maintenance tasks he has little resource left over to pursue the Semantic Web/LOD vision he has long been an advocate for at his institution. (Related to scenario 7.3.6 and scenario 7.3.7)

7.4. Scenarios

7.4.1. Becka's company wants to create a U.S. Civil War research database. The licensing job is done: the company she works for has applied Named Entity Recognition technologies to a massive cache of Civil War-era letters and documents, extracting and disambiguating names

with a high degree of precision. They have also acquired a substantial database of all regimental records from the era. The remaining job is to align these two, matching extracted names to regimental entries. (Related to persona 7.3.1)

7.4.2. Much of the content in the institutional repository Christina is working on predates the advent of ORCID identifiers. In some cases, the researchers responsible for publications in the repository have never had ORCID identifiers; in others, they subsequently acquired them after publishing in the repository; and in other cases, researchers may have assigned themselves more than one ORCID identifier. Christina really just needs to sort this mess out – given a list of academic staff records, she needs to supply the relevant ORCID identifiers where these exist. (Related to persona 7.3.2)

7.4.3. Darren's institution has acquired an archive of letters and documents relating to a prominent early 20th-century local author – or rather, several local authors, as the writer emerged from a thriving local literary scene, which has been well-documented by various academic members of staff. These academic staff members have supplied him with a detailed, but highly nonstandard, 'authority list' of individuals mentioned, not all of whom meet the bar of 'literary warrant'. Darren would like to associate these various names with standard identifiers, in cases where such identifiers exist; where they don't, he plans to mint his own, and would like to register and make these publicly available to others. (Related to persona 7.3.3)

7.4.4. Joris's institution has large and (mostly) well-maintained authority lists for artists, artefact typologies, and for the subjects of representational art. He would like to align these with other relevant vocabularies, such as the Getty vocabularies and ICONCLASS. In cases where there's no match, he would like to add artists to Wikidata and/or feed them back to the Getty. (Related to persona 7.3.4)

7.4.5. Nathalie has assembled an extensive collection of materials related to case studies in human rights violations. This collection is global in scope and multilingual in character, including public-domain materials from the Nuremberg trials, legal documents related to the Yugoslav civil wars, and archives of oral testimony witnessing the Rwanda conflict of the early '90s. The material is well-cataloged and curated, according to prevailing local standards and practices, but these are not necessarily all commensurate with each other. There are accordingly a number of areas Nathalie wants to align, here: legal documents with a legal taxonomy; named individuals with UN and national archival databases; places with geo-identifiers. But she is aware of the highly charged political and emotional content of the material – and of the corresponding risk of offence or legal action in the case of misidentification of individuals. (Related to persona 7.3.5)

7.4.6. Nuno wants to be able to federate online catalog search across all Spanish university libraries. In the first instance, this means aligning records at the level of the Item (in FRBR-speak), so that users can track down precise editions of particular texts. In an ideal world, however, Nuno would like to operate at the level of Works, aligning such records with this level of abstraction in e.g., VIAF. (Related to persona 7.3.6)

7.4.7. Nuno has recently discovered that The European Library has produced an alignment of the subject headings used by the Library of Congress, the Bibliothèque Nationale, and the German National Library. Nuno would like to add the headings used by the Spanish national library into the mix as well. (Related to persona 7.3.6)

7.5. User Stories

7.5.1. As a … user familiar with controlled vocabulary concepts
I want to … search a universe of name authority records that may or may not match the name of a person on an item being described
So that … add a globally unique identifier to a field in a descriptive metadata record.

7.5.2. As a … system administrator
I want to … add a source of local authority records to be harvested
So that … name authority records from another source can be added/reconciled against the database.

7.5.3. As an … aggregator who receives metadata in various languages
I want to … optionally specify the language of the submitted name-value in a standards-compliant fashion
So that ... the number of false matches is minimised and that languages with complex nominal-derivation rules can potentially be catered for in a specialised manner.

7.5.4. As a … developer for an aggregator
I want to … be able to configure the rules used to generate matches
So that … I can optimise these matches in the way that best fits my particular use-case

7.5.5. As a … developer for an aggregator
I want to … have the rules and logic used to determine a match and Its confidence-rating reported to me in a clear and understandable way
So that … I can make an informed judgement regarding which aspects of the match-algorithm pipeline are working well, and which need to be tweaked or eliminated.

7.5.6. As a … production manager for an aggregator
I want to … deploy the RaaS stack locally
So that … my production processes are minimally dependent on outside infrastructure, and so that I can customise the RaaS stack if required.

7.5.7. As a … developer for an aggregator
I want to … to develop extensions/plugins for the reconciliation service without having an in-depth awareness of the RaaS codebase internals
So that … I can customize the functionality of the service to suit my particular use-case

7.5.8. As a … developer for an aggregator

I want to ... be able to develop extensions or plugins for the reconciliation service without in-depth awareness of RaaS codebase internals
So that ... I can focus my attention on my particular reconciliation issues, rather than on the tooling itself.

7.5.9. As a … vocabulary curator
I want to … find additional data to add to the canonical version within my vocabulary
So that … the vocabulary I curate is as extensive as possible

7.5.10. As a … cataloger
I want to … enter a name and an attribute (or number of attributes), and get an exact match, or only have to sort through a few records to get an exact match
So that … I can get a unique identifier for my entity, and efficiently determine if I need to create an authority record.

7.5.11. As a … metadata professional
I want to … have a place to definitively state that one entity is X and is not Y
So that … we do not have to duplicate effort to make these assertions again.

7.5.12. As a … metadata professional
I want to … align my local authority lists with external vocabularies and contribute to the external vocabularies
So that … we can publish our authorities.

7.6. Select Work

The need for reconciliation services is clearly one felt by the research and library community as a whole, as evidenced by a convergence of effort by multiple independent bodies in this area.

Notable other work currently being undertaken in this area, and exhibiting strong overlaps with the concerns of the IMLS Sharable Local Authorities forum are:

- The LD4 Working Group on Reconciliation: https://github.com/LD4/ld4-community-recon
- The Pelagios Linked Pasts Working Group (http://commons.pelagios.org/groups/linked-pasts/), which identified a reconciliation service as central to its aim of linking historical data sets into a unified graph representation
- SNAC (http://snaccooperative.org/) offers to assist with a member project's matching and then adding new entities into SNAC.
- VIAF Project (https://viaf.org/), which is taking various national name authority files and creating a reconciled list.
- Europeana (https://www.europeana.eu/) is beginning to work on service to assist sources of data to help data needs (as of July 2017)

- Authify, a RESTful reconciliation module that offers several search and detection services (component of the SHARE-VDE project: http://share-vde.org/)
- Karma (http://usc-isi-i2.github.io/karma/) is an information integration tool that enables users to quickly and easily integrate data from a variety of data sources including databases, spreadsheets, delimited text files, XML, JSON, KML and Web APIs. Users integrate information by modeling it according to an ontology of their choice using a graphical user interface that automates much of the process. Karma learns to recognize the mapping of data to ontology classes and then uses the ontology to propose a model that ties together these classes. Users then interact with the system to adjust the automatically generated model. During this process, users can transform the data as needed to normalize data expressed in different formats and to restructure it. Once the model is complete, users can published the integrated data as RDF or store it in a database.
- The FOLIO Project (https://www.folio.org/) is an open source library services platform currently under initial development. One of the apps anticipated for this platform is an authority service. As a platform based on RESTful network communication principles, other applications would be able to make use of the authority service in the FOLIO platform.

7.7. Recommendations for Future Work

To bring the concept of a Reconciliation-as-a-Service offering to fruition, SLNA-NF participants anticipate several research and development needs.

7.7.1. Recommendations for publishing name authority data from local systems

Local name authority data can take many forms in various systems. For instance, an integrated library system may store locally defined names in a MARC 790 field, an institutional repository may use a MODS <name> field with a custom ‘authority' attribute, and an archival finding aid system may store the locally defined names in a specific relational database table. The variety of these storage mechanisms makes interoperability difficult. A common data model and interchange format are needed to make this data available to reconciliation services. The published data needs not only the local identifier and authority string; it also needs the context in which the name authority is used (such as institutional affiliations, subject headings or topical keywords of published material, and dates).

7.7.2. Mechanisms for registering and discovering local published name authority data

The inherently local nature of the name authority data created in an institution's installation of various systems adds significant challenges to the discovery of that data by reconciliation systems. A registry or other discovery mechanism is needed for operators of RaaS servers to discover local name authority data sets for harvesting. Defining a template for local sources of authority data is in scope so registry entries can be machine-parsed by software such as the RaaS below. Local sources of authority data should download or upload metadata about its own data

source or desired external data sources from/to a clearinghouse (similar to datahub.io, with the goal of sharing metadata rather than the dataset itself). Both clearinghouse and local sources can use VoID to describe the linked datasets. The clearinghouse shall enable data curators to share, update, and discover authoritative data sources':

- namespace
- provider and contact information for error report
- software/tool and version
- data model and version
- available web services and version
- available data classes
- data retention policy

7.7.3. Software to aggregate, preprocess, cluster, and display name authority records

The first two needs listed above are prerequisites for any RaaS provider. With those two needs satisfied, any number of organizations or services providers can offer RaaS services to specific communities. The SLNA-NF participants also desired a prototype, reference implementation, or open source software project that is the RaaS itself. In discussing this software, the SLNA-NF participants noted that transparency in the clustering algorithm is needed so the user understands why matches or non-matches occur, including the weighting of terms in making that determination. Human intervention is seen as a necessity in the clustering/matching process; managing expectations of what will be needed for the human work is important. Ideally, it should be possible to adapt the algorithm based on the improvements suggested through human intervention.

Two types of reconciliation are needed: a batch/retrospective interface for reconciling bulk metadata that exists now and an interactive interface with drop-downs and type-ahead features that optimize the user experience for selecting the appropriate local name authority record. It would be useful to separate the user interface from the back-end database searching so that such an authority service can be used by any type of software -- catalog, institutional repository, finding aids, etc. Supporting OpenRefine's reconciliation service in batch mode would be very useful, as well.

7.7.4. Clearinghouse for matching algorithms

In addition to RaaS software, several SLNA-NF participants desired a mechanism to catalog and describe matching algorithms used by various projects to cluster name entities. Absent a common machine-oriented way to describe these algorithms, the clearinghouse can contain code examples with narrative text.

The institution that supports the clearinghouse might also be the forum to address the organizational challenges and governance issues in the authority sharing ecosystem. One example is the separation of identifier management system from the multiple aggregating systems to facilitate the persistent identifier registration of local authorities.

8. Conclusion

There has been and will continue to be a need for local systems and services to balance the need of local name authorities with more robust standards, registries, and aggregators. The scalability of this will necessitate a look at the linked data structure to keep synchronization in check as well as limit duplication of efforts. Efforts of aggregators have shared the practical learning experiences of working with library authority data. The SLNA-NF was able to leverage that knowledge to determine some of the next steps needed towards the consensus need for a Reconciliation as a Service product.

The library community faces a paradigm shift regarding our approach to authority and entity data creation. The availability of alternative approaches in very traditional models is becoming a reality. For instance, the Program for Cooperative Cataloging (PCC) is undergoing a pilot to test ISNI as a NACO alternative throughout 2018.[32] If successful, this has the potential to lower the threshold for many institutions to contribute entity data. Meanwhile, the PCC URIs in MARC task group[33] produced recommendations for inclusion of subfields $0s and $1s to store URIs in MARC bibliographic data, bringing aspects of linked data into MARC workflows; these recommendations were approved by the MARC Advisory Committee and are being implemented as of early-2018. As institutions consider incremental steps to a productionizing linked data, more institutions are considering methods of incorporating a broader diversity of authority and entity data sources into their metadata practice.

With a focus centered on minimum viable specifications, data provider obligations, workflows and reconciliation-as-a-service, the SLNA-NF surfaced a number of issues related to the benefits and blockers encountered by attempting to share local name authority data. Much more work is needed to find solutions to the issues raised through this forum. The community gathering that transpired during SLNA-NF and in other venues will be the necessary mechanism as institutions look to more diverse authority practices.

[32] Program for Cooperative Cataloging (2017/2018). *ISNI Pilot*. https://wiki.duraspace.org/x/m5s2BQ
[33] Program for Cooperative Cataloging (n.d.). *PCC Task Group on URIs in MARC*: https://www.loc.gov/aba/pcc/bibframe/TaskGroups/URI-TaskGroup.html

9. Appendixes

9.1. 2016 October 24-25 In-person Meeting Agenda

2016 October 24

08:30 - 9:00	Coffee
09:00 - 09:30	Welcome (Chew Chiat Naun) Introduction and goal-setting
09:30 - 11:00	Lightning Talks 1 - Joan Cobb: ULAN: From Batch Contribution to LOD - Janifer Gatenby & Andrew MacEwan: ISNI use cases at national libraries - Daniel Pitti: Archival context and authority control Facilitator: Jason Kovari Note-takers: Isabel Quintana, Chew Chiat Naun
11:00 – 11:15	Break
11:15 - 12:30	Lightning Talks 2 - Timothy Hill: Things in Strings: reification and enrichment of Europeana metadata - Tiziana Possemato: Local authority data in Casalini's Share platform - Jean Godby: OCLC entities Facilitator: Nancy Fallgren Note-takers: Nancy Lorimer, Anna Neatrour
12:30-13:45	Lunch
13:45 – 14:45	Wishing Out Loud session Facilitator: Chew Chiat Naun Note-takers: Hannah Tarver, Nancy Fallgren

14:45 - 16:00	Lightning Talks 3 - Jing Wang: Authority control in research ecosystem - Hannah Tarver: Authority control at UNT Libraries - Ryan Wick: Opaque Namespace - Jason Kovari: Entity Management at Cornell : Plans Facilitator: Anna Neatrour Note-takers: Isabel Quintana, Chew Chiat Naun
16:00 - 16:15	Break
16:15-17:30	Panel Discussion - Review of day's discussions - Goals, format for Day 2 Panelists: Jason Kovari, Carl Stahmer, Michelle Durocher, Suzanne Pilsk Note-takers: Nancy Lorimer, Chew Chiat Naun
17:30	Adjourn
18:30	Dinner

2016 October 25

09:00 – 12:15	Plenary Discussions, Topics (selected): - What is sharing (value and purpose)? - Understanding data consumers - Minimum viable product specifications - Matching and identity resolution - Models, policies & workflows Facilitators: Jason Kovari, Chew Chiat Naun
12:15 - 13:30	Lunch
13:30 - 13:40	Nettie Lagace: NISO and Community Engagement

13:40 – 15:00	Wrap-up and Next Steps - Reporting - Outreach - Discussion going forward - Action items Panelists to include: Diane Hillmann, Jean Godby, Nancy Fallgren, Steven Folsom, Nettie Lagace, Corey Harper Note-takers: Hannah Tarver, Anna Neatrour
15:00	Adjourn

9.2. 2017 April 10-11 In-person Meeting Agenda

2017 April 10

09:00 - 09:30	Arrival & Coffee
09:30 - 10:00	Introductions & meeting scope
10:00 - 10:30	SNAC demo - Worthy Martin and Daniel Pitti (University of Virginia)
10:30 - 10:45	Break
10:45 - 12:00	Institutional & Practice : Lightning Talks (in-depth) + Discussion - Anna Neatrour (University of Utah) - Mark Phillips (University of North Texas) - Peter Murray (Index Data) - Facilitator: Jason Kovari (Cornell University)
12:00 - 13:00	Lunch (on own)
13:00 - 14:15	(Inter)national & Providers : Lightning Talks (in-depth) + Discussion - Simeon Warner (Cornell University) - Michelle Durocher (Harvard University) - Paul Frank (Library of Congress) - Facilitator: Jason Kovari (Cornell University)
14:15 - 14:30	Break
14:30 - 15:00	IPFS - Matt Zumwalt (Protocol Labs, Inc.)
15:00 - 15:15	SHARE-VDE - Michele Casalini and Tiziana Possemato (Casalini Libri)
15:15 - 15:25	Getty Vocabularies : Update - Joan Cobb (Getty Research Institute)
15:25 - 16:00	Survey scoping - Facilitator: Chew Chiat Naun (Cornell University)

16:00 - 16:30	Wrap-up day
16:30	Adjourn for day

2017 April 11

09:00 - 09:15	Arrival & Coffee
09:15 - 10:30	Reconciliation as a service - Facilitators: Timothy Hill (Europeana) and Peter Murray (Index Data)
10:30 - 10:45	Break
10:45 - 12:00	Minimum Viable Product Specification - Facilitators: Anna Neatrour (University of Utah) and Chew Chiat Naun (Cornell University) - Speaker: Isabel Quintana (Harvard University)
12:00 - 13:00	Lunch (on own)
13:00 - 14:15	Data Provider Obligation - Facilitators: Janifer Gatenby (OCLC Leiden) and Jean Godby (OCLC)
14:15 - 14:45	Outreach & community engagement - Facilitators: Nettie Lagace (NISO) and Daniel Pitti (University of Virginia)
14:45 - 15:00	Break
15:00 - 15:30	Define / identify follow-on projects
15:30 - 16:00	Wrap-up & Next Steps
16:00	Adjourn for Day

9.3. Workflow Questions

Workflow gathering : issues to consider

The following issues to address in gathering information about workflows currently in production; these do not have to be prescriptive. The IMLS SLNA-NF has a separate group planning to conduct a survey (in extension of existing work); this effort is much more directed at individual institutions and also purely focused on workflows.

1. Do you re-use existing authorities? Which sources? (e.g. NACO, VIAF, SNAC…)

2. Which sources do you use for gathering data? (e.g. local, professional organization, telephone or to publisher / author, wikipedia….)

3. What are the factors that made you decide to create a new authority?

4. Do you contribute to a collective file? (NACO, VIAF, SNAC, ISNI…) How often? How? (online, batch)? Formats used?

5. Maintenance: how often do you perform maintenance or QA/QC testing?

6. Do you do periodic sampling for quality? Do you have specific queries for this?, e.g. searching for unique source or unique name for verification

7. How do you store and/or implement authorities? Do you manage data locally or via a (cloud) service or via APIs? What data formats do you use: XML, RDF, Linked Data, etc.?

8. Do you conduct reconciliation on your authorities? If so, what is the process around reconciliation? what specific problems are addressed, fixed, checked, etc in the reconciliation process?

9. If you find an error, where do you correct it? How do you diffuse the correction?

10. Do you use any other tools for maintenance, e.g. synchronization reports, notifications from a collective file

11. Do you use crowdsourcing for maintenance and enrichment?

12. What are your steps in creating a new authority?

13. For each stage of the workflow, how many staff are dedicated to a given stage, what are the staff roles?

14. Do you partner with 3rd parties or with other local parties to perform any steps of your workflow (e.g. local IT or Academic Technology department, CrossRef, VIAF, etc.)

15. Do you maintain documentation of your workflow, if so, is it freely available?

9.4. Draft Survey Instrument (not conducted)

As part of the second in-person SLNA-NF meeting, participants decided to create a survey instrument to better understand authority practice at institutions. The aim of this survey is to discover practices related to local identity data, motivations for sharing, and barriers for sharing this data more broadly. Time constraints prevented the group from finalizing the question set or conducting this survey during the IMLS grant period. We provide these questions in case other communities would like to pursue this work.

Survey questions:

What are your main sources of identity data?

What data or entity types do they cover?

- Multiple choice options: Persons; Organizations; Creative Works or Publications; Geographical entities, places; Subjects; Other

To which shared sources do you contribute directly?

What characterizes a good experience when sharing data with aggregated or community sources?

Which of your services do these sources support?

- Multiple choice options: Library catalog; Institutional repository; Research output tracking; Digital collections; Other special project (please give details); Other

In what cases do you create data outside those shared systems (e.g.: LCNAF, ULAN)?

- Multiple choice options: Entity types not supported by the system; Local data not supported by the system; Unverified or incomplete data; Training of staffing accreditation barriers; Process efficiency or data flow considerations; Other

How would you describe the current sharing or distribution model for the system in which you create authorities?

- Multiple choice options: Strictly local use only; Local creation of data feeding into aggregation; Shared platform with domain or regional partner; Reuse of data captured from other sources; Other (please specify)

Are there projects for which you would like to create identity data but currently do not? If so, what are the barriers?

Which of the following best describes your role within your organization?