Report for PCC Task Group on the Creation and Function of Name Authorities in a Non-MARC Environment

April 5, 2013

Charge to task group

1. Think broadly and practically about identities (personal, corporate and family) in both an RDA and a linked data environment and how they function within it. What will that environment look like? What are the key conceptual differences from the current authority record environment?
2. Identify the key changes that are needed to current authority record systems, structures, and guidelines to support the new environment, including the impact on OCLC/NACO normalization rules. [http://www.loc.gov/aba/pcc/naco/normrule-2.html]
3. Prepare a report to describe (as specifically as possible) those key changes, problem areas and proposed solutions. What are the barriers to moving forward, and what can PCC do to eliminate them? What can NACO catalogers do now to move in this direction? [http://www.loc.gov/aba/pcc/rda/RDA%20Task%20groups%20and%20charges/Non-MARC-Name-Authorities-TG-charge.docx]

Introduction: Identity in RDA

Identity for persons, corporate bodies, and families in RDA is defined by a set of elements beginning with the preferred form of the entity's name. Some of these elements may be combined with the preferred form of the name to create an authorized access point for the entity. Other elements may be recorded as information about the entity but not as part of the authorized access point. In some cases an element may be both recorded separately and used as a part of the authorized access point.

The range of additions to the preferred name that may be used to form a unique authorized access point is not entirely settled. Under LC-PCC guidelines, adding dates is expected; adding a fuller form of name is allowed to break a conflict, and also in cases where a cataloguer considers it useful for identification. Important to the work of this task group, the Joint Steering Committee (JSC) recently approved changes to extend the range of qualifiers allowed for personal names under RDA (cf. 6JSC/BL/3/Sec final, http://www.rda-jsc.org/docs/6JSC-BL-3-rev-Sec-final-rev.pdf), and 6JSC/BL/4/Sec final, http://www.rda-jsc.org/docs/6JSC-BL-4-Sec-final.pdf).The new instructions will be incorporated into RDA with the 2013 update. LC-PCC has not yet provided a policy statement on the new instructions.

Identifiers are included among the elements associated with persons, corporate bodies, and families ("named entities"). At 9.18.1.1, 10.9.1.1, and 11.12.1.1, RDA states that, "The identifier serves to differentiate that [named entity] from other [named entities of the same type]." Identifiers can be used both as key data for retrieving more complete information about an entity and as a representation of the entity for use in declaring that entity's relationship to another entity, such as a work, expression, manifestation, item, another person, another corporate body, etc. In this way an authorized access point and an identifier have generally been seen as parallel representations of the entity.

Authorized access points in RDA are created by combining a name with a defined set of informative qualifiers when available. There are other possible approaches to differentiating one entity name from another. In some systems (the Internet Movie Database, or IMDb, for example) a system-supplied arbitrary numbering, "I, II, III, IV, …", is used to differentiate entities with the same name but known to be different. An identifier associated with a named entity's authorized description could be appended to an authorized access point to perform a similar differentiating function when such an identifier exists. However, it has not been practice under AACR2 or RDA to use identifiers as part of the differentiating information in an authorized access point. The meaning of such identifier data would likely be mysterious to users if they encountered it in displays.

Under AACR2 and LCRI the allowable differentiating qualifiers that could be added to a personal name heading were limited. In some cases this meant that a person could be differentiated, but the person's name heading could not. To cope with this situation, undifferentiated personal name authorities were created in which specially annotated 670 fields were clustered to indicate different identities established under an undifferentiated personal name heading. In contrast, undifferentiated name authorities were never created for corporate bodies under AACR2/LCRI practice. Authorities for families were created at a very general level to serve as broad subject headings following LC's Subject Headings Manual, H1631.

This practice of creating undifferentiated personal name authorities could continue under RDA. Authorized access points are preferably unique in RDA, but it is not required. RDA 8.11 defines an "undifferentiated name indicator" as "a categorization indicating that the core elements recorded are insufficient to differentiate between two or more persons with the same name." The associated LC-PCC PS states that the undifferentiated name indicator is "a core element for LC/PCC."

Despite this longstanding practice, there are also practical reasons to question the continued use of or need for undifferentiated name authorities. The ambiguity of which identity an undifferentiated authority's identifier refers to undermines its usefulness as an identifier. Moreover, the instability of the individual and multiple identities which that identifier may have referred to over time make the value of such records highly debatable.

As context for the current task group's work, PCC stated in the charge:

> Following significant input at a public forum in June 2012 at ALA Anaheim, the PCC Policy Committee decided that
>
> - PCC does not want to proliferate undifferentiated personal name records in the LC-PCC Name Authority File going forward;

- There is no urgency to split apart undifferentiated records that already exist, but we do need to define new practices to achieve differentiation and implement them (PoCo will begin work soon on guidelines);
- A project to split apart undifferentiated records in the retrospective file would best be accomplished after we have experience with those new practices;
- We need to change our paradigm from name headings/records defined by unique text strings to unique identities, and to define a path from one to the other; and
- We will form a Task Group to define that path, to identify the issues and barriers to achieving it, and make recommendations for next steps to move us toward this new paradigm.

This task group report addresses two areas of concern to PCC as reflected in the charge and the context statement. Part One considers the options for discontinuing the creation of undifferentiated personal name authorities and for eliminating this type of authority from the retrospective file. Part Two is a more general exploration of the role of name authorities and identity data generally in a post-MARC, linked data environment.

Part 1

Alternatives to Undifferentiated Personal Name Authorities

The task group discussed three basic options to replace the current use of undifferentiated personal name authorities.

Option 1. Use the unique LCCN identifier alone to differentiate the persons represented by authorities. 100 fields would no longer have to be unique, and the LC/NACO Heading Comparison rules would no longer be needed.

Option 2. Revise the RDA instructions and/or LC-PCC PS for 9.18 to include the LCCN identifier among the allowable additions to an authorized access point when needed to break a conflict.

Option 3. Revise the RDA instructions and/or LC-PCC PSs for Chapter 9 to ensure that informative qualifiers can always be formulated to differentiate one personal name authorized access point from another.

Discussion

Option 1 has the virtue of simplicity. All LC/NACO Authority File records have an LCCN identifier (LCCN in turn is the basis for a URI at id.loc.gov), which guarantees that any person who can be differentiated can have a differentiated authority record. The identifier could be hidden from users, but could still serve to differentiate one name access point from another. Systems would be free to devise solutions for how best to differentiate the display of such names

for users, e.g., by supplying arbitrary local numbering (I, II, III, …) or by displaying one or more of the recorded elements from the authority record in addition to the name. The need for the cataloger to select additions for the authorized access point could be replaced by system-driven concatenations of data elements with the preferred name.

On the negative side, Option 1 would likely introduce significant variability in how persons would be represented to users in different contexts. If the primary point of differentiation is seen only by the system and users are shown either local arbitrary numbering or a varying mix of qualifying data, it may be more difficult for them to recognize the same person in different contexts. While sufficient in principle for functions the system needs to perform, it may not be sufficient for supporting the user tasks that RDA is based on.

It will also require significant reconfigurations of current system programming. While there was general consensus in the task group that identifier-based differentiation is the optimal approach to managing identity data, there was also concern that relying on identifiers alone at this point to distinguish otherwise identical authorized access points is not yet practical. It remains an important goal for the future and will be discussed further in the Part 2 of the report.

Option 2 is similar to Option 1, but it displays the differentiating identifier as part of the authorized access point when needed. This would preserve the notion of an authorized access point shared across many systems as an aid to users seeking to find and identify persons as they work across different sources.

However, the identifier element is not currently authorized by RDA for inclusion in an authorized access point. Moreover, by itself an identifier provides little useful information for a user seeking to identify and select a named entity from a set of similar name access points. The LCCN could also be ambiguous in an international context. Lastly, the LCCN identifier is available as a distinguishing element only for names that are established. Option 2 offers no solution to the problem of differentiating unestablished name access points on bibliographic records. Being able to achieve differentiation even in the absence of a corresponding authority will be of value to PCC libraries and others in a shared cataloging context.

Option 3 essentially extends the current practice of differentiating authorized access points by the addition of informative qualifiers. By increasing the range of allowable qualifiers to embrace such narrowly specifying phrases as "Author of Title ABC," "Subject of photograph [URL]," etc., PCC could ensure that all named persons could be assigned a unique authorized access point. Since this approach is essentially an extension of existing practices, it would have the smallest impact on system programming. Since such phrases are typically used to differentiate identities in existing undifferentiated personal name authorities, Option 3 would also offer an approach to resolving the retrospective problem.

On the other hand, implementing Option 3 would make the sorting order of authorized access points for personal names highly unpredictable. One advantage of the limits placed on allowable

qualifiers in the past has been that alpha-numeric sorting was more predictable, for example, clustering many common names under an ordered set of dates. However, this sorting simplicity has already been compromised under AACR2/LCRI by the use of even a limited range of qualifiers to resolve conflicts, and would continue to be so under RDA. With a wider range of qualifiers, there is also increased risk that multiple authorities may be created for the same person simply because an existing authority was not seen or not recognized. However, the fact that the authorities are differentiated means that the authority records can contain discretely recorded RDA elements that can also be made searchable and aid in finding and recognizing existing authorities.

A more subtle risk of Option 3 is that it has the potential to focus attention on unique text strings rather than on identifiers. The task group is unanimous in asserting that over time using identifiers to declare the relationships between named entities and resources and their relationships with other named entities is the preferred course. Options 1 and 2 have the advantage of keeping identifiers more in focus as primary information keys.

Lastly, the task group recognizes that certainty around questions of identity differentiation is not always possible. If the only evidence for a name is, for example, two documents in the same discipline but on different topics, and no other relevant information can be found, there may be no way to be certain whether the name represents two persons or one person. Nevertheless, creating an undifferentiated authority is not a desirable solution, if for no other reason than such a record with its 670s organized in separate identity clusters implies cataloger's judgment that two people are involved. If cataloger's judgment determines that there are two persons, they should be established on two separate authorities. If cataloger's judgment opts for there being only one author, one unique name authority citing both titles should be created.

If either of these instances of cataloger's judgment is found later to be in error, the error can be corrected. Merging two authorities into one is relatively straightforward and does not significantly disrupt the connection between identity and identifier(s). Creating a single authority when two were needed is a more serious error, and may obscure the need to add new information to the authority. Once the error is recognized, separating two identities which have been mistaken for a single person presents no more challenge than separating the identities linked to undifferentiated authorities currently does. The proposal later in this document for deconstructing existing undifferentiated personal name authorities could also be a model for how to handle unique personal name authorities which are later found to represent two persons ambiguously. The need for a way to mark personal name authorities as undifferentiated which are found to have erroneously identified two or more persons with a single 100 field and LCCN in the correction process could justify a new use of the 008/32=b code in the future after the intentional creation of undifferentiated authorities has been discontinued.

Recommendations

The task group recommends Option 3 to address the problems of undifferentiated personal name authorities in the current LC/NACO Authority File and in existing systems and Option 1 as the preferred solution when systems are able to support it.

1. Revise LC-PCC PS 9.19.1.1, Differentiating Authorized Access Points for Persons, to authorize use of descriptive qualifying phrases with preferred names as permitted under RDA (cf. 6JSC/BL/4/Sec final, http://www.rda-jsc.org/docs/6JSC-BL-4-Sec-final.pdf), including the kind of phrases currently used to label 670 clusters on undifferentiated authorities. This will eliminate the need for creating undifferentiated name authorities in the future.
2. Implement the use of identifiers rather than access point text strings as the primary match point between bibliographic authorized access points and their authority records. This step will depend on PCC's determination that system functionality to support identifier-based matching and entity representation is in place.
3. Request that LC revise DCM Z1 008/32 to discontinue the practice of creating and adding to undifferentiated personal name authority records, and of revising them to become differentiated authorities.

Deconstructing Existing Undifferentiated Name Authorities

The LC/NACO Authority File currently contains close to 60,000 undifferentiated personal name authority records. The clustering of 670 fields on the record indicates the different identities represented by the undifferentiated access point, e.g.,

100 1 $a Young, Frank A.

670 $a [Co-editor of The use of ceramics in surgical implants]

670 $a Internatl. Biomaterials Sym., 1st, Clemson Univ., 1969. $b The use of ceramics ... 1978 (a.e.) t.p. (Frank A. Young)

670 $a [Author of Duluth's ship canal and aerial bridge ...]

670 $a nuc86-99754: His Duluth's ship canal and aerial bridge ... c1977 $b (hdg. on MnHi rept.: Young, Frank A.; usage: Frank A. Young)

The convention of beginning each identity cluster with a bracketed identifying phrase offers an option for automated deconstruction of most undifferentiated name authorities. If these phrases are regarded as allowable qualifying information which can be added to a preferred name access

point, a program could be written to divide all of the existing undifferentiated name authorities into separate authorized access points on separate name authority records.

One question raised in the task group was how to code "last resort" qualifying phrases such as "Author of Duluth's ship canal and aerial bridge ..." Currently subfield $c is defined in MARC21 to include "other words or phrases associated with a name." The contents of subfield $c can be reflected in other, more precisely designated fields. A phrase such as "Author of Title ABC" could be coded in 368 $c under current MARC21 definitions, which would ensure a place for that phrase if the form of the heading is revised to something more conventional in the future. Titles themselves can be recorded in field 672, newly approved for "Titles related to the Entity Represented by the Authority Record." (http://www.loc.gov/marc/marbi/2013/2013-01.html) In the examples below modeling the new practice, we have used subfield $c. (Use of 368 $c is permitted by DCM Z1, 368 and by the LC/NACO guidelines to the MARC21 Authority Format.)

Some undifferentiated name authorities also include 675 fields for sources which could provide no information. In some cases, the sources consulted bear a relationship to specific cluster identities but not to all the identities on the record. Nevertheless, for the sake of simplicity such 675 fields could be copied to all the generated differentiated authorities. Similarly, all the generated authorities should carry the 667 field called for by current LC procedures (DCM Z1, 008/32). The resulting machine-generated authority might look in part like this:

008/32 = a

100 1 $a Young, Frank A. $c (Author of Duluth's ship canal and aerial bridge …)

368 $c Author of Duluth's ship canal and aerial bridge …

667 $a Formerly on undifferentiated name record: [LCCN of undifferentiated name record]

670 $a nuc86-99754: His Duluth's ship canal and aerial bridge ... c1977 $b (hdg. on MnHi rept.: Young, Frank A.; usage: Frank A. Young)

The DCM Z1, 008/32 section also instructs that when a single identity is left on an undifferentiated name authority after others have been moved to their own authorities, the undifferentiated authority should be recoded as a unique personal name authority (008/32=a). Given the uncertainty this creates over time regarding the relationship between the authority record ID and which person it represents, this practice should be discontinued. Identifiers are intended to be shared with other systems and to be used as consistent, reliable keys for referring to an established entity. Changing the entity to which an identifier refers in the source system is bad data management practice, especially in an environment where sharing of identifiers across systems is increasing.

A preferable practice will be to establish all the identities currently found on undifferentiated name authorities on new authority records. The existing undifferentiated authority records should be "decommissioned," which might be accomplished in several ways; but they should not be removed from the file, which ideally should contain a record of each instance of the past practice for use when looking up an authority record's status and history by LCCN. Possible recoding for the decommissioned authorities:

LDR/05 = c (Record status: Corrected)

008/09 = b (Kind of record: Untraced reference)

008/32 = b (Undifferentiated personal name)

008/14 = b (Heading use - main or added entry: not appropriate)

008/15 = b (Heading use – subject added entry: not appropriate)

008/32 = b (Undifferentiated personal name: retain the existing code)

100 1 $a Young, Frank A.

The task group considered calling for the addition of a standard annotation in a 100 $c subfield text, e.g., $c (Undifferentiated person), to ensure that the name form itself is not blocked from being used in the future for a properly differentiated identity. On the other hand, such typically unqualified common names have already proven problematic in cataloging. With ready recourse to additional qualifiers, the task group has opted to recommend that no such qualifier be added. Preserving the undifferentiated name heading in its original form will ensure that the persons who share a preferred name will get distinguishing qualifiers.

As with any large data set, there are anomalous cases which will present challenges to automated processing. The task group recognizes this and recognizes that some replacing of undifferentiated name authorities will require manual intervention. We also recognize that some of the automatically generated new authorities will be duplicates of existing authorities, given the LC/NACO community's imperfect record of removing the 670 identity clusters from undifferentiated authorities when establishing them separately. These concerns notwithstanding, the task group calls for an automated project to deconstruct the existing undifferentiated personal name authorities with confidence that the resulting problems will be more readily solvable than the problems currently presented by the undifferentiated authorities.

Additional information regarding the conversion of existing undifferentiated authorities is provided in the Appendix to this report.

Recommendations for resolving existing name authorities

1. Request that LC undertake a project with appropriate partners to perform automated conversion of existing undifferentiated personal name authorities into separate, differentiated authorities for each identity. Volunteer PCC support for this project.
2. Request that LC revise the existing undifferentiated name authorities after the new, separate authorities are generated to mark the undifferentiated authority records as untraced reference records (008/09=b) not to be used for authorizing headings.

Part 2

Name Authorities in a Non-MARC Environment

Identity in a linked data environment

Linked data prescribes that entities and relationship types be represented by resolvable URIs (whenever possible) in a simple subject-predicate-object syntax. While much recent attention has been focused on the publishing of bibliographic and authority data in linked data formats, this is only one aspect of adopting a linked data model. Equally if not more important is the inclusion of relationships with external data sources within one's own data. It is integrating external data with one's own more than simply exposing one's own data that, when engaged in by all parties, will enable more fluid navigation and leveraging of the richness of the linked data environment for enhanced discovery.

URIs for certain kinds of data structures (e.g., identity records) can function as web-level IDs for the entities those data structures define and describe. Such URIs--e.g., the id.loc.gov URI for a named entity and the VIAF URI for the same entity—can serve as identifiers for that entity in a variety of contexts, provided the URI is stable in form and has a fixed relationship to the entity. Each URI can have different data and services associated with it. For example, for Barbara B. Tillett:

http://id.loc.gov/authorities/names/n88102106 provides access some elements of the authority and to other formats for the LCCN n88102106 data, including MARC/XML, MADS/XML, SKOS, JSON, and RDF.

http://viaf.org/viaf/77390479 provides access to data about Dr. Tillett from many major authority files in a variety of format options.

Library of Congress and OCLC have already begun projects which incorporate these linked data URIs into external contexts, e.g., adding VIAF URIs to id.loc.gov data and to Wikipedia pages, and adding Wikipedia page URIs as part of VIAF data. The availability of these machine-actionable links expands the kinds of data services, navigation, and retrieval that can be offered to users.

Library name authority data has high value particularly in relation to the named entities which function as the creators, contributors, and subjects of monographs and serials. It can provide identifying information and unique IDs for many persons and other named entities and can do so in the open web as shown above, not only in library systems. Efforts to expose library name authority data at the network level are already well underway. However, the linked data environment is much larger than library catalogs, and the resources which libraries provide to their users contain many names not found in library authority files. Many of those names have or will have linked data identity records in external systems, like the International Standard Name Identifier (ISNI) registry, the Open Researcher/Contributor Identity registry (ORCID), and others. This presents many challenges to libraries, including:

- How to relate external named entity identifiers and information to library authority data;
- How to relate external named entity identifiers and information to library bibliographic data;
- How to structure library authority data to optimize its utility in a linked data environment as well as in library systems.

Features of the new environment

A major reason for moving away from MARC is to enable greater interoperability with data in other systems which follow more broadly based standards, such as XML and RDF. More important than the coding used is the set of objects, relationships, and categories which a community intends to represent. RDA has defined aspects of named entities which have had high value for finding, identifying, and selecting such entities. The success of an encoding for RDA name authorities is measured by its ability to express these elements and their relationships clearly and unambiguously. That expression should have both machine-actionable and human-friendly forms. The former ensures greater integration of library data in networked applications. The latter ensures that the data has value for end users.

Interoperability means interdependence. Library catalogs have a long history of working across interdependent files. Bibliographic records incorporate name headings for entities whose definition and metadata management occurs in name authority files; likewise for subjects and some titles. The work of representing entities and concepts is distributed across a variety of files and data structures. The management of linkages between these files has tended to be based on shared text strings representing names, concepts, etc., but here too libraries already have a history in particular applications and particular communities of using identifiers to forge more reliable and durable links between records.

The linked data environment extends the possibilities for this kind of interoperation. An ISNI URI can be embedded in a catalog record to provide more than just a name for a person. The

ISNI provides a reliable unique identifier for that person linked to additional metadata about the person. The ISNI can also be coordinated with other representations of a person, such as one found in the LC/NACO Authority File; but such correspondence is not necessary for it to serve as a unique identifier with access to useful metadata. This opens up possibilities for drawing on a larger pool of identity records for named entities than catalogs have used up to this point.

It is possible that automated ways could be found to derive LC/NACO Authority File records for persons and other named entities from files like the ISNI registry, but such an approach could present major maintenance challenges and limited extensibility. An alternative approach would be to enlarge the pool of acceptable sources of named entity metadata records for use in bibliographic descriptions. Given the inevitable ambiguity of name text strings, such an approach would have to depend on URI identifiers as the primary keys for specifying named entities. This would be a significant change for many automated library systems to implement, but its potential for enhanced services, improved data management, and greater integration between library-managed data and the larger networked environment makes it worth the investment.

When authoritative identity is distributed across multiple systems, it becomes important to manage the relationships between entity representations in those systems cooperatively. The inclusion of related IDs in each system's record data, e.g., of national authority record IDs in VIAF clusters and of VIAF cluster IDs in id.loc.gov records, is one way of managing such relationships. The British Library is working to include a large batch of ISNI IDs in LC/NACO Authority File records as part of its work with the ISNI registry. Alternatively, open tables devoted to expressing and maintaining same-as and other specified relationships between named entity IDs could be built for this purpose.

The mechanisms by which such relationships are determined will need a high degree of reliability. As the work on VIAF has found, there is a balance that must be struck between matching a larger number of authorities from source files and matching those authorities correctly. Automated algorithms to determine matches are essential to the task of determining matches across large, disparate identity data sets, but algorithms will inevitably match together some authorities for different persons and fail to match some authorities for the same person. A viable approach to the task of determining the entity relationships between files will need both algorithmic tools for high-volume matching and scalable methods for manual additions and corrections to such matching.

Encoding to convey the source and other transactional and contextual aspects of the named entity's representation—for example, the most recent check of any embedded data against the source, or "pending" status information from the source, or aspects of the relationship between the entity and the resource—would likely also be needed. Something like this is already being modeled in the BibFrame Initiative's "authority" data structure, which is designed not as a traditional authority but as a connector between two separate metadata descriptions. "[BibFrame]

Authorities are not designed to compete or replace existing authority efforts but rather provide a common, light weight abstraction layer over various different Web based authority efforts to make them even more effective." (http://www.loc.gov/marc/transition/pdf/marcld-report-11-21-2012.pdf)

Enabling the use of a wider range of identity data sources in catalog records would make the mix of data available about each entity more variable, since different identity systems record different kinds of data. However, uniformity of content has never been a hallmark of traditional library name authority files, which have been built over time in accordance with an evolving set of standards and with limited efforts at retrospective maintenance. In itself, the lack of uniformity should not be a significant obstacle to employing multiple sources of authority. It will present the same challenges that traditional files do for systems that seek to incorporate metadata about named entities from different systems into a unified index for discovery and retrieval.

Differing definitions of what constitutes named entities will be more of a problem. Identity systems which disagree about whether or not pseudonyms are separate entities or whether corporate bodies with a history of name changes are a single entity or several entities will not have one-to-one mappings to each other. Either the mappings expressed between them will be less precise, or more precise and nuanced expressions of relationships will present added complexity for users.

Changes needed to authority record systems and structures

Some have questioned whether catalog data needs to move beyond the "record" into a more fluid model of statements joined in varying combinations based on the linked data URIs they share. In an environment of interdependent entity representations the idea of a fixed, authoritative record may be problematic. For example, if an LC/NACO Authority File record has a declared relationship to entity representations in other systems such as ISNI and ORCID and updates to the information in those systems has an impact on how the LC/NACO authority record is presented in different contexts, the "authoritative" form of the record might be hard to pin down. However, it is unlikely that the need for complex, structured representations of what is known about a named entity compounded of multiple discrete statements will ever go away. The term "authority record" in this report refers primarily to the latter concept of a structured set of semantically defined elements and relationships, including identifiers which function both as collocation points and as conduits for information from external sources.

Traditional authority control focuses on expressing identity by means of unique headings which name discrete entities. In card catalogs, arranging text strings alphabetically was the only available technology for collocating entries. The need for unique headings was driven by the fact that only the heading itself typically appeared when an authorized entity was being referenced. The unique name heading for a person was what ensured that bibliographic records which cited that person as a creator or contributor or subject could be gathered under one heading. Variants of the name were also important. By recording variant forms of name in authorities as

alternative access points, catalogers could direct users from those name variants to the one chosen for collocation in library catalogs.

While the management of identity via preferred and variant access points was well understood, the need to describe the named entity in the authority was not immediately recognized. Authorities from early in the history of the LC/NACO Authority File sometimes contain nothing but the established heading and an abbreviated citation of a title in LC's catalog. It was assumed that the cited bibliographic record could carry much of the burden of more clearly identifying the person named, e.g., what subjects an author wrote about, where the author published, with what co-authors, and so on.

As authority records and files were shared, authority practice moved toward providing more complete information in the authority record itself, making the authority more valuable as a stand-alone representation of the named entity. This trend is reflected in the additional designated data elements which RDA has defined for describing persons and corporate entities. However, there is still a tension between the model of authority control which focuses on name access points and the model which focuses on describing and categorizing named entities. Further work is needed on the data structures for authorities to enable the recording of both qualifier terms grammatically appropriate for use as access point qualifiers and controlled terms which place the entity in relation to other entities and in relation to other categorical terms (e.g., relating a person in the specific category of "Pediatricians" to the general category of "Doctors").

As discussed in Part 1 of this report, a transition from an access-point-focused model of authorities to one focused on entity description will depend on shifting the work of designation and differentiation of named entities from the established access point to a unique identifier for the person. In principle, once this shift has been achieved, the need for access points to be unique could be de-emphasized. Systems should be able to construct differentiated representations of named entities based on the identifier, the set of data elements describing the entity in the authority, and the relations the named entity has to objects in the library's collection.

Different systems may still choose to provide specified forms of an entity's name for various purposes. Some systems will provide personal name forms in direct order for ease of readability. Some will provide names in an inverted order more suited to conventional alphabetical indexing, or formulated according to non-English naming conventions. Some may choose to require a form of the name which is unique within the system, i.e., something very like the LC/NACO Authority File's unique authorized access points. In any case, the effort to include one or more forms of an entity's name should be driven by user needs rather than simply by the weight of past practice. It should not be assumed that all sources of identity data will see a need to provide a unique pre-coordinated version of each entity's name.

What should be assumed is a commitment to the rule that unique identifiers represent single entities as the system defines them. Resolving the issue of undifferentiated name authorities in the LC/NACO Authority File is essential for meeting this requirement.

Discovery systems will also need to devise better ways to support queries about named entities. As the range of data specified in library authorities and available via links to external identity data sources widens, systems will be better able to construct search result displays which focus on the facets of named entities and not just those of bibliographic objects. Users should be able to define classes of persons whose work is of interest, e.g., 18th century Russians, or women mathematicians, or composers of film music. Innovation is needed in systems that support technical services tasks as well. Catalogers should be able to see faceted aspects of persons with common names to assist with the task of selecting the right authority for an item in hand. Systems will need to interoperate with external ID registries and data sources to harvest and combine data from other systems to extend and enhance the metadata provided about an entity.

In a non-MARC, linked data environment, expectations will be higher that each authority represents a single entity. Differentiation is not just a problem for personal name access.

- Subject entities can be undifferentiated (e.g., LC does not differentiate geographic features with the same name below the county level).
- Past and current practice has tended to combine the representation of a work and its original language expression, two separate entities under FRBR, in a single authority, making the coding of elements defined only for one entity or the other problematic, e.g., Form of Work (Work), Content Type (Expression), and Associated Language (Expression).
- Differentiation rules for family names in a shared metadata environment are still in flux.
- Conventional practices around geographic names which have changed over time present a different kind of ambiguity. For example, the LC/NACO Authority File record for Sri Lanka stands for both that country since 1972 and as a subject, for that place for its entire history.

Ensuring that library authority records of all types correspond specifically to concepts and objects as the larger community defines them will increase the value of the authority record identifiers and the metadata they reference.

Paths forward

Part 2 of this report is much more speculative and open ended than Part 1, so its recommendations are not specific courses of action but suggested focus points for additional exploration and development.

- Develop policies and practices to express links between LC/NACO Authority File records and identity records in other systems following linked data principles.

- Consider developing policies, coding, and practices to enable the use of registered IDs outside the LC/NACO Authority File in bibliographic descriptions.
- Engage other sectors of the information environment—system developers, service providers, ID registries, cultural heritage institutions, etc.—in exploring the use of URIs and linked data syntax for expressing and managing identity metadata
- Model and promote the use of faceted searching and results display for entity metadata derived from authorities in library discovery and data management systems.
- Take a lead role in reconfiguring the relationship between library metadata and metadata drawn from other sources and in realigning expectations regarding cooperation and collaboration across sectors in the information community.
- Consider developing tools and techniques outside the LC/NACO Authority File for expressing relationships between identified entities and between relationship categories found in different systems.

Specific questions for LC/PCC debate

Should a full record in the LC/NACO Authority File be required for authorized access points in BIBCO records, or do IDs from other sources of registered name identity (Wikipedia, VIAF, other library authority files, ORCID, ISNI, etc.) have a comparable place in BIBCO cataloging?

What is PCC's role in asserting and revising relationships between LC/NACO Authority File records and identities in other registries?

If an LC/NACO Authority File record references other identity records, what metadata from those records (if any) should become part of the LC/NACO Authority File record?

PCC Task Group on the Creation and Function of Name Authorities in a Non-MARC Environment

Kevin Ford

Stephen Hearn (chair)

Thom Hickey

John Hostage

Mary Mastraccio

Richard Moore

Berit Nelson

Beth Picknally Camden

Gary Strawn (consultant)

Jessalyn Zoom

Appendix 1: Revising existing undifferentiated personal name authorities

The task group recommends a project be undertaken by LC and PCC to revise existing undifferentiated personal name authorities in the LC/NACO Authority File. In support of that suggestion and with the help of Gary Strawn, this appendix provides some additional detail about how such a project should be scoped and carried out.

The LC/NACO Authority File contains nearly 60,000 undifferentiated personal name authority records—too many to revise efficiently by hand. An examination of the records in the file (i.e., authority records coded 008/32=b) indicates that many of the records can be successfully converted programmatically. For this work, a process and a number of default values and field tests must be defined.

The proposed process takes each authority coded 008/32=b, test it for various exception cases, and if no exceptions are found, generates new authorities coded 008/32=a for the identifiable persons listed in the existing authority and recodes the existing authority as an "untraced reference" record, 008/09=b. The exception cases are reported out for manual processing.

In a programmatic evaluation of 59,552 undifferentiated name authorities by Gary Strawn, 95% of the records were able to be converted. Counts for the unconverted exception cases in this test were:

2439 records with improper configuration of 670 fields

126 records with $t title subfield

117 records with 500 fields present

15 records with only one bracketed 670 present

These preliminary results indicate that most of the undifferentiated name authorities currently in the LC/NACO Authority File could be programmatically converted to generate unique, separate authorities with at least minimal identifying information. The remainder of the undifferentiated authorities could be the focus of a project managed by LC and PCC. The project would distribute the remaining records and monitor the progress of manual conversion by volunteer NACO institutions. Undifferentiated authorities containing non-roman 400 fields should be directed to NACO libraries with JACKPHY language skills.

The list of proposed exception cases below is longer than that used in the tests just described, but further evaluation may find ways to automate significant portions of the exceptions casework.

Proposed defaults for the newly generated authorities

008/32=a (Undifferentiated personal name=Differentiated)

010=new LCCN

100=name from 100 plus parenthetical subfield $c containing bracketed descriptive phrase that begins a cluster of 670 fields related to the same person

667 Formerly on undifferentiated name record: [LCCN of undifferentiated authority]

670s associated with identity described in leading bracketed 670

675 repeats any 675 on all generated differentiated authorities

Proposed defaults for the revised undifferentiated authority

008/09=b (Kind of record=Untraced reference)

008/14-008/16=bbb (not appropriate for use as a heading)

008/32=b (Undifferentiated personal name=Undifferentiated)

100 = no change

666 $a Names formerly established under the undifferentiated name heading [former 100] have been established separately.

667 $a DO NOT USE, undifferentiated personal name authority

667 noting request for additional information (retain on undifferentiated authority)

667 noting "Record covers additional persons" (delete from undifferentiated authority)

Proposed exceptions to automated processing

008/32=b authorities with 400 fields that are more specific than the 100

008/32=b authorities with non-roman 400 fields

008/32=b authorities with 500 fields

008/32=b authorities with 663, 664, 665, and 666 fields

008/32=b authorities with no bracketed 670s

008/32=b authorities with consecutive bracketed 670s

008/32=b authorities with 100 containing subfield $c

008/32=b authorities with 100 containing $t

008/32=b authorities with JACKPHY data in a 400 field

Sample converted records

ORIGINAL UNDIFFERENTIATED RECORD

Rec stat: c Entered: 800709 Char: a
Type: z Upd status: a Enc lvl: n Source:
Roman: ? Ref status: n Mod rec: Name use: a
Govt agn: ? Auth status: a Subj: a Subj use: a
Series: n Auth/ref: a Geo subd: n Ser use: b
Ser num: n **Name:** **b** Subdiv tp: n Rules: c

001 777911
005 20110812141833.0
010: : |a n 79089972
040: : |a DLC |b eng |c DLC |d OCoLC
100:1 : |a Hutton, Peter
670: : |a [Author of Guide to Java]
670: : |a His Guide to Java, 1974: |b t.p. (Peter Hutton)
670: : |a [Illustrator of Motorcycles]
670: : |a Cave, R. Motorcycles, 1982 (a.e.) |b t.p. (Peter Hutton)
670: : |a Phone call to Gloucester Press, 7/20/82 |b (Peter Hutton, a British, no other info. available)
675: : |a NUC, 1978-81.

REVISED UNDIFFERENTIATED RECORD

Rec stat: c Entered: 800709 Char: a
Type: z Upd status: a Enc lvl: n Source:
Roman: ? Ref status: n Mod rec: Name use: b
Govt agn: ? Auth status: a Subj: a Subj use: b
Series: n Auth/ref: b Geo subd: n Ser use: b
Ser num: n **Name:** **b** Subdiv tp: n Rules: c

001 777911
005 20110812141833.0

010: : |a n 79089972
040: : |a DLC |b eng |c DLC |d OCoLC |d XXX
100:1 : |a Hutton, Peter
666: : |a Names formerly established under the undifferentiated name heading Hutton, Peter have been established separately.
667: : |a DO NOT USE, undifferentiated personal name authority

NEW DIFFERENTIATED RECORD 1 OF 2

Rec stat: n Entered: 130315 Char: a
Type: z Upd status: a Enc lvl: n Source: d
Roman: ? Ref status: n Mod rec: Name use: a
Govt agn: ? Auth status: a Subj: a Subj use: a
Series: n Auth/ref: a Geo subd: n Ser use: b
Ser num: n **Name:** **a** Subdiv tp: n Rules: c

005 20130315103355.0
010: : |a [new LCCN]
040: : |a DLC |b eng |c DLC
100:1 : |a Hutton, Peter **|c (Author of Guide to Java)**
667: : |a Formerly on undifferentiated name record n 79089972
670: : |a His Guide to Java, 1974: |b t.p. (Peter Hutton)
675: : |a NUC, 1978-81.

NEW DIFFERENTIATED RECORD 2 OF 2

Rec stat: n Entered: 130315 Char: a
Type: z Upd status: a Enc lvl: n Source: d
Roman: ? Ref status: n Mod rec: Name use: a
Govt agn: ? Auth status: a Subj: a Subj use: a
Series: n Auth/ref: a Geo subd: n Ser use: b
Ser num: n **Name:** **a** Subdiv tp: n Rules: c

005 20130315103355.0
010: : |a [new LCCN]
040: : |a DLC |b eng |c DLC
100:1 : |a Hutton, Peter **|c (Illustrator of Motorcycles)**
667: : |a Formerly on undifferentiated name record n 79089972
670: : |a Cave, R. Motorcycles, 1982 (a.e.) |b t.p. (Peter Hutton)
670: : |a Phone call to Gloucester Press, 7/20/82 |b (Peter Hutton, a British, no other info. available)
675: : |a NUC, 1978-81.

Appendix 2: Glossary of Acronyms and Initialisms

AACR2 – Anglo-American Cataloguing Rules, 2nd edition.

http://www.aacr2.org

http://en.wikipedia.org/wiki/Anglo-American_Cataloguing_Rules

BibFrame Initiative – Bibliographic Framework Transition Initiative

http://www.loc.gov/marc/transition

http://www.bibframe.org

DCM Z1 – Library of Congress Descriptive Cataloging Manual, Z1: Name and Series Authorities

http://www.loc.gov/catdir/cpso/dcmz1.pdf

ISNI – International Standard Name Identifier

http://www.isni.org

JACKPHY languages – Japanese, Arabic, Chinese, Korean, Persian, Hebrew, Yiddish

http://www.loc.gov/loc/lcib/0712/cataloging.html

JSON – JavaScript Object Notation

http://www.json.org

LCCN – Library of Congress Control Number

http://www.loc.gov/marc/lccn.html

LC-PCC PS – Library of Congress-Program for Cooperative Cataloging Policy Statements

http://www.loc.gov/aba/rda/lcps_access.html

LCRI – Library of Congress Rule Interpretations

http://www.loc.gov/cds/products/product.php?productID=43

MADS – Metadata Authority Description Schema

http://www.loc.gov/standards/mads/

MARC – Machine Readable Cataloging

http://www.loc.gov/marc/

ORCID – Open Researcher and Contributor ID

http://www.orcid.org

RDA – Resource Description and Access

http://www.rdatoolkit.org/

http://www.rda-jsc.org/rda.html

http://en.wikipedia.org/wiki/Resource_Description_and_Access

RDF – Resource Description Framework

http://www.w3.org/RDF/

SKOS – Simple Knowledge Organization System

http://www.w3.org/2004/02/skos

URI – Uniform Resource Identifier

http://en.wikipedia.org/wiki/Uniform_resource_identifier

VIAF – Virtual International Authority File

http://www.viaf.org

XML – eXtensible Mark-up Language

http://www.w3.org/XML/