Crossing Silos: Assessing the Utility of Identity Attributes in Name Reconciliation

Ruth Kitchin Tillman and Gala Campos Oaxaca

abstract: Even as librarians have spent the past twenty years documenting the rich world of scholarly communication beyond the catalog, repositories and catalogs too often remain completely siloed from each other. Current practices and tools to unite the two focus entirely on matching names, an imprecise method requiring substantial time spent on review. This article presents results of an experiment incorporating the attribute and citation data present in Library of Congress Name Authority Records and local faculty database records into the process of authority reconciliation. Adding tests for employer affiliation, educational history, and academic department produced improved, highly accurate match results.

Introduction

Authority work serves "to ensure consistency … so that [the user] has to search under one and only one heading to find records associated with names, subjects, and other access points."[1] Name authorities collate all works by a particular author, even if published under variations of the same name, and disambiguate these from works by other authors, even those with identical names. In addition to authorized and variant name forms authorities may contain datapoints, referred to here as "identity attributes," which support author disambiguation and contextualize an author's work. Rules for authority work in library catalogs have existed since at least the 1940s and national and international databases publish name authorities for reuse across institutions.[2]

However, while authorities are often well-integrated into a library's catalog systems, those designing institutional repositories and faculty profiling systems have ignored or struggled to implement such structures. Challenges range from the scale of articles published to the wide variety of ways in which publishers represent a person's name.[3] Systems showcasing institutional research often handle local disambiguation by using

portal: Libraries and the Academy, Vol. 25, No. 2 (2025), pp. 299–319.

institutional directory identifiers. Without a way of linking institutional identifiers to national authorities, there is no way to programmatically connect authors' works in institutional repositories (IRs) with their works in the library's catalog. The result is an incomplete picture of institutional research and creative accomplishment.

The disconnect between authorities leaves academic librarians unable to answer questions related to the representation of faculty work in their collections. For example, it would be difficult for most to determine which works in their catalog were created by faculty in their school's African American Studies program without performing a manual search for each faculty member. This slows the work of collection assessment and development. It may also damage the library's reputation within their institution when librarians are unable to answer such questions about their own collections.

Over the past decade, librarians and technologists developed a set of tools and practices for pairing representations of a person or thing in two different systems, a process often referred to as "reconciliation." These tools enable projects connecting institutional authors with their national authorities at greater speed and scale than entirely manual searches. Unfortunately, most of these authority files only support reconciliation based on matching names with other names. When present, identity attributes are consulted during the review process but have not been incorporated into standard methods of querying.

This research inquiry emerged from the gap between existing processes and datapoints that they leave unused. Is it feasible to incorporate identity attributes into a programmatic reconciliation between local author data and the Library of Congress Name Authority File (LCNAF)? How does one address the differences between such datasets? What is the accuracy of resulting reconciliation attempts, and can this approach lower the time spent on manual review?

Literature Review

The challenge of reconciling local name data with external authorities is shared by those working across institutionally-specific collections from analog and digital archives to institutional repositories. Name reconciliation projects are generally undertaken to support linked data implementation by augmenting local data with external identifiers or to improve data quality through alignment with widely-established authorities, particularly the Library of Congress's Name Authority File.[4] Although such work may be outsourced to vendors, performed locally using the popular OpenRefine reconciliation service, or achieved through a combination of the two, it always relies on quantifying the similarities between two text strings, generally a local name field and the main or variant labels of the authority.[5]

Yet more data exists. Authority records often contain identity attributes that support author disambiguation and contextualize an author's work. Library of Congress Name Authority Records (NARs), for example, may contain datapoints describing a person's occupation, institutional affiliation, and field of activity. But while OpenRefine supports adding values from additional columns in the reconciliation process, these datapoints are not encoded in the LCNAF, the Virtual International Authority File (VIAF), and many other library name authorities in a way that can be used during reconciliation and thus cannot be queried. Instead, these are often manually reviewed in a time-consuming secondary phase of the project. Jeremy Myntti and Anna Neatrour recount that after Backstage Library Works processed a 2012 export of all names and subjects from the Willard Library's CONTENTdm repository, an intern "spent approximately 100 hours over the course of 10 weeks" reviewing the resulting reports.[6] In an analysis of an extract of 1,000 names from Duke's repository, Moira Downey reports that "it quickly became apparent that the extracted text strings alone would be insufficient information to confidently disambiguate the author represented by the string from among the numerous candidate identities returned by the Application Programming Interface (API) without recourse to examining the individual publication with which that entity was associated," a process for which there were "insufficient resources."[7]

The authors found no evidence in the literature of other attempts to reconcile local and Library of Congress name authorities which make programmatic use of identity attributes during the reconciliation process. This is likely due to the complexity of the Library of Congress's linked data syntax. For example, Bria Parker and Adam Gray described the structure of the JSON-LD authority records as "challenging to parse" programmatically and abandoned their use.[8] This article demonstrates possible solutions to these challenges and undertakes to advance the practice of name authority reconciliation.

Project Description and Goals

As at many institutions, the Pennsylvania State (Penn State) University's institutional repository and library catalog exist in entirely separate metadata silos. Authors in the institutional repository are disambiguated using their institutional directory identifier, whereas those in the catalog are represented by their Library of Congress Name Authorities Records. Any project to reconcile these two through traditional means would have required a great deal of time spent manually reviewing potential matches, and the institution lacked an appropriate place to record results.

The launch of Penn State's Researcher Metadata Database (RMD) faculty profiles in 2020 provided an opportunity for a new approach to reconciliation. Faculty profiles are both a source of data about a faculty member and a place for the resulting reconciliation to be recorded. Once recorded, a pairing between faculty profile and external authority could be used for future enhancements to either the local catalog or repository.

The initial phase of this project consisted of time-consuming experimentation with data review and testing processes. Using a set of authorities with known affiliations, the team spent significant time determining the most efficient methods to query appropriate fields in each record and account for any variations in encoding. The resulting Python script identified NARs with close name matches and conducted additional tests based on the following local data:

- current institutional affiliation,
- occupation,
- educational history, and
- department name.

The sections that follow describe the process of data preparation and script development, assess the utility of identity attributes in matching records, and provide recommendations for methods of replicating the work elsewhere.

Data Sources and Working Datasets

Researcher Metadata Database Background

The Researcher Metadata Database is a locally developed, centralized clearinghouse for metadata about research being performed at Penn State. The metadata is organized into profiles, primarily representing members of the Penn State faculty. Unlike VIVO, ORCID, or Elsevier Pure, which publish public-facing faculty profile webpages, RMD's data is primarily accessed via an API. Querying this API with a person's Penn State directory ID returns JSON-formatted data representing all or part of their RMD record.

RMD data is harvested from internal and external sources, harmonized for each person, and organized into sections which represent major sections of a CV. These include department affiliations, educational history, peer-reviewed publications, other publications, sponsored research, presentations, and advising history. A substantial portion of this information comes from ActivityInsight, software that Penn State faculty are encouraged to use to support tenure and promotion activities.

The completeness of an RMD profile depends on data completeness in internal and external systems. For example, someone who reached the rank of full professor before the system was implemented may not use ActivityInsight at all or only track the recent work which must be documented in their 5-year post-tenure reviews. Others represented in RMD may not have ActivityInsight accounts due to the nature of their positions. External sources may be similarly incomplete or not published in a way that can be indexed into the system. Therefore, some profiles are blank except for the person's directory information: name, email, title, department, and phone number. A detailed description of RMD's datapoints and their sources can be found on the project's Github README.[9]

Creating a Working Faculty Dataset

In this project, the team focused on the subset of researchers holding faculty appointments at Penn State. The manager of the Faculty Activity Management Services Team provided the researchers with a spreadsheet listing only faculty in the system. This spreadsheet included each person's directory ID, first name, family name, and faculty status. The team then downloaded profile objects for each person using their directory ID and the API.

Because the RMD profile object is intended to generate a CV-like output, each person's entire name is only represented in a single "name" field. This would have been difficult to compare with authorized name forms in LCNAF, which are in a "Family name, first name" order. Fortunately, the spreadsheet included the first name and family name from the Penn State directory as distinct fields. The team augmented the download script to add additional fields for first name, family name, and an inverted form made by combining the two.

The resulting JSON objects contained only the fields that would be used in this project:

- Penn State directory ID
- first name
- family name
- inverted name
- name
- affiliation (department, center, or organization within the university, occasionally null)
- educational history (sometimes null)

Library of Congress Name Authority File Background

Name Authority Records are created by catalogers at the Library of Congress and those participating in the Program for Cooperative Cataloging's NACO program.[10] Name authorities connect all works by a particular author and disambiguate these from works by other authors, even those with identical names. In most cases, they are created when the cataloger's library has acquired a work for which the person is primarily responsible (author, editor, composer) and the person is not yet represented by a name authority record.

Best practices and available fields for NAR creation change over time. Newer records are likely to include the many post-Resource Description and Access (RDA) fields which provide context about the person's work and support disambiguation. While catalogers may choose to update older records, there is no program for systematic updates other than the removal of deprecated fields. Some NARs, therefore, consist of only a person's authorized name and citations to the sources the cataloger used in its creation. Others include robust description, such as fields describing a person's occupation, field of activity, and organizations with which they had been affiliated.

Creating a Working Authority File Dataset

There are currently no APIs that provide access to the additional data points of a LCNAF record. The team worked directly from a local copy of the LCNAF, which can be downloaded in bulk as MADS/RDF (Metadata Authority Description Schema in RDF) from https://id.loc.gov/authorities/names.html. JSON-LD was selected because of Python's strength in parsing JSON.

Storing the data on a local computer allowed (and required) the team to develop search strategies from scratch. The first steps toward developing such a search were challenging. While the RMD JSON objects were simple enough for a beginner to parse, the MADS/RDF JSON-LD syntax of the LCNAF download posed a major challenge. [11] The team began by identifying a set of records that contained most or all of the desired fields. They conducted initial experiments to identify and extract each field, working in batches that were small enough for easy manual review. During this process, the team solidified an understanding of the MADS/RDF JSON-LD syntax and identified several variations that might occur, such as when a record contains a descriptive term that is not part of a Library of Congress vocabulary.

While the research team used a computer with significantly more processing power than those normally distributed to library employees, the size of the LC Name Authority

File as JSON-LD posed a challenge. The original file, downloaded on October 16, 2023, was 42GB. Because each record was on a separate line, the team used a simple bash script to split the file into 47 separate files of 250,000 records (lines) each, with a remainder of about 150,000 in the final file.

These smaller files were still 546MB to 1.1GB each. The team ran a Python script to transform full authority records into minimal records containing only textual values of fields that would be used for matching. This script also removed records with the types madsrdf:CorporateName, madsrdf:FamilyName, and madsrdf:NameTitle, leaving records which had only the type madsrdf:PersonalName. The resulting files, while still large, now contained 112,814 to 190,474 records and ranged in size from 54-120MB.

All records contained:

- id: URI as /authorities/names/[identifier]
- authorized_name: single element that has @type of both madsrdf:Authority and madsrdf:PersonalName, value of the madsrdf:authoritativeLabel (MARC 100)
- citation_data: list of madsrdf:citationSource / MARC 670a and madsrdf:citationNote / MARC 670b statements.

Many records contained some of the following optional elements:

- alt_names: any elements whose @type is both madsrdf:PersonalName and madsrdf:Variant, value of any madsrdf:variantLabel (MARC 400), formatted as a list.
- activity: values of any madsrdf:fieldOfActivity statements (MARC 372a), formatted as a list
- occupations: values of any madsrdf:occupation statements (MARC 374a), formatted as list
- organizations: values of any madsrdf:organization statements (MARC 373a) found within elements of type madsrdf:Affiliation, formatted as a list.

Testing Process

The team wrote the testing process to take a tiered approach. First, the script checked the working record against each LCNAF record to determine whether its authorized or alternative name forms were a 95 percent match or higher for several variant forms of the person's name. For records in which the script found at least one such match, it conducted four additional context tests against:

- affiliation with Penn State or the university's name in the record citation,
- an occupation containing the word "teacher," "faculty member", or "librarian,"
- affiliation with the person's educational institution(s), and
- a field of activity related to the person's institutional department or the presence of the department name in the citation note.

The first two tests used the same static set of values for each member of the faculty: affiliation with Penn State and a set of occupations. The second two dynamic tests used values from the faculty member's RMD profile: the names of their educational institution and department. Because RMD's educational institution names are entered by faculty members in the internal ActivityInsight software and NAR citations are uncontrolled text that catalogers entered to provide context, neither field can be relied on to contain standardized institutional names. These tests employed a variety of methods to look for matches. Any records with contextual matches were added to the resulting spreadsheet output. This section provides a brief overview of each test. See Appendix A for a pair of example records and testing walkthrough.

Testing Names

The first set of tests ran between names created from the RMD file and the authorized and alternative names in the Library of Congress (LC) records. Dates, the most common form of additional information in an LC name, were stripped. Because a person may express their name in a variety of ways in different contexts, the team used all available data in the RMD objects to create up to three variants of the faculty member's name:

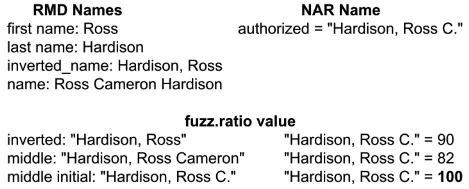
- Last, First
- Last, First Middle (if present in RMD)
- Last, First Middle Initial (derived from middle or from directory ID if the middle letter was not "x")[12]

The script calculated Levenshtein distance, the same method used by OpenRefine's reconciliation, to evaluate the degree of match between each pair of names tested. It used the FuzzyWuzzy library for Python, which returns the evaluation as an integer. After testing a set of variations, created from 80 faculty names against their known authorized forms, the team set a minimum match threshold for the project at 95. While this would not capture all known matches, it also limited the number of irrelevant names the project would return.

After testing each name variation, if the highest match integer returned remained below 95, the script would move on to the next LCNAF record. If the highest match integer was 95 or higher, the script would run additional tests described in the following sections, write the record to the output file, and move on to the next LCNAF record (see Figure 1).

Static Test: Institutional Affiliation

The team searched for institutional affiliation using two fields: affiliation data and citation data. Affiliation data may include a person's employers and educational institutions (see Figure 2). While the field uses a controlled value list of organizations established as corporate names in the LCNAF, there was no single authority representing the university. Corporate name authorities exist for many Penn State colleges and campuses and a person's affiliation might be recorded as "Penn State Harrisburg" or "Pennsylvania State University. Department of Meteorology" rather than "Pennsylvania State University." After research, the team determined that a search checking for substrings "Penn State" or "Pennsylvania State" would match all university-affiliated organizations. If this test found a match, the script recorded it and moved on to the next test. If not, it tried again in citation data.



RMD Names
first name: Ross
last name: Hardison
inverted_name: Hardison, Ross
name: Ross Cameron Hardison

NAR Name
authorized = "Hardison, Ross C."

fuzz.ratio value

inverted: "Hardison, Ross"	"Hardison, Ross C." = 90
middle: "Hardison, Ross Cameron"	"Hardison, Ross C." = 82
middle initial: "Hardison, Ross C."	"Hardison, Ross C." = **100**

Figure 1. Example of calculating the fuzz.ratio for name variations.

Citation data, uncontrolled text entered by catalogers at the time of record creation or update to record the sources of information, is a much older MARC field and might contain information related to the author's affiliation. Because the field is free-form and was originally printed on catalog cards, a cataloger might choose to abbreviate or represent the institution's name in a greater variety of ways. Based on review of citation data in the sample of 80 matches, the team identified the following possibilities: "Pennsylvania State," "Penn State," "PennState," "Penn. State," and even "State College, PA" (the location of Penn State's largest campus).

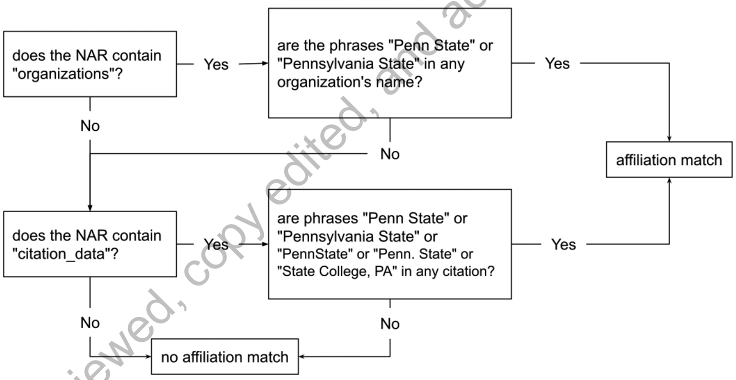

does the NAR contain "organizations"?
Yes
are the phrases "Penn State" or "Pennsylvania State" in any organization's name?
Yes
No
No
affiliation match
does the NAR contain "citation_data"?
Yes
are phrases "Penn State" or "Pennsylvania State" or "PennState" or "Penn. State" or "State College, PA" in any citation?
Yes
No
No
no affiliation match

Figure 2. Testing organizations and citation data for Penn State affiliation.

Static Test: Occupation

Authority records may contain occupational data, generally a controlled field from a relevant Library of Congress vocabulary. The Library of Congress Demographic Group Terms (LCDGT) introduced the term "University and college faculty members," which might be used on any faculty member's record. However, the occupation field predates this vocabulary. It may contain other terms, most often from older controlled vocabularies. Moreover, the cataloger creating an NAR may not include information about a person's faculty role at all, depending on the context in which the NAR was created.

The team initially identified the Library of Congress Subject Heading (LCSH) "College teachers" which may describe anyone from adjunct lecturers to distinguished professors at colleges or universities, as the most relevant term used prior to the publication of LCDGT. However, subsequent data assessments revealed that catalogers had sometimes used more specific LCSH terms, such as "Philosophy teachers." A search for the term "teacher" as a string somewhere in the field returned all kinds of teachers, most of whom worked at the college and university level. Because Penn State's librarians are members of its faculty, the team added the occupation "Librarian." While research faculty are another type of Penn State faculty, the team was unable to identify a unified term that appeared suitable to identify them.

The final version of this test, shown in Figure 3, looked for the presence of any of the following strings in an authority record's occupation field as string:

- "teacher,"
- "University and college faculty members," or
- "Librarian."

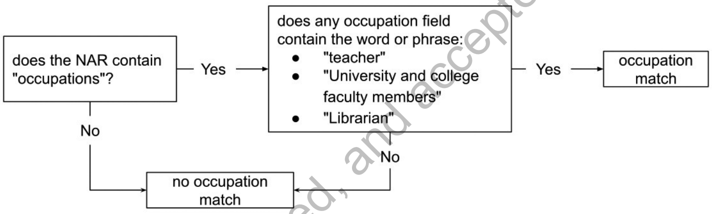

does the NAR contain "occupations"?
Yes
does any occupation field contain the word or phrase:
• "teacher"
• "University and college faculty members"
• "Librarian"
Yes
occupation match
No
No
no occupation match

Figure 3. Testing occupations field for a faculty-related occupation.

Dynamic Test: Educational Institution

Many RMD records contained information about a person's educational history extracted from ActivityInsight. If this data existed, the script checked it against the NAR's organizational affiliation and citation data. Because both RMD educational data and NAR citation data are completely uncontrolled, the team identified a common set of variants for university names and wrote a data cleanup and variant creation process into the script. Variants were based on the most common abbreviations found in both RMD and the NAR. Only universities were transformed, as data related to advanced degrees was more often recorded in the NAR.

First, the script cleaned up any abbreviations or non-standard practices employed by the faculty member. For example, if the faculty member had entered an institution's name ending with a space and capital "U," the script would add a variation in which " U" was replaced with " University." Similarly, "U of" was replaced with "University of." Any initial "The" was removed. These changes ensured that the data was more likely to match a controlled form of organizational affiliation.

The script then created a set of abbreviations for use in searching citations: "Univ," "Uni," and "U." It generated these abbreviations with and without a period, "Uni" and "Uni.", as both forms were found in citations. Table 1 outlines the steps in cleanup and variation generation. For the sake of visual clarity, it does not include variants with periods.

Table 1.

Example of name variations created for two university names as entered by faculty

Process Step	Sample 1	Sample 2
Starting data	The University of Florida	Cornell University
Cleanup: Initial "The" removed	University of Florida	n/a
Cleanup: "U of" or final "U" replaced with "University"	n/a	n/a
Version to test with organization listing	University of Florida	Cornell University
Variant Generation: "University" to "U," "Univ," and "Uni"	U of Florida, Univ of Florida, Uni of Florida	Cornell U, Cornell Univ, Cornell Uni
Full list to test with Citation	The University of Florida, University of Florida, U of Florida, Univ of Florida, Uni of Florida	Cornell University, Cornell U, Cornell Univ, Cornell Uni

Once the script had generated these names, it tested the cleaned-up version against the NAR's organizational affiliation and, if no match was found, tested each of the variants against citation fields using the same process as the static affiliation test above (Figure 2). When a positive result for one institution was found, the script then moved on to test any additional institutions listed in the RMD educational history, as multiple institutional matches increase overall confidence in a record match.

Dynamic Test: Department

The final test sought a match between a person's departmental affiliation in their RMD record and the activity and citation fields of the NAR. After the team examined a large data sample, it was determined that the best method of matching activity and department would be to test whether any value from the NAR's field of activity was found in the name of a department.

For example, if one value in an NAR's field of activity list was "Astrophysics" and the RMD record being tested had the department field "Astronomy and Astrophysics," the affiliation match would be recorded as "true." Because the values in field of activity might come from Library of Congress Subject Headings, they might be subdivided, as in "Latin America--History." If the script found a "--" it would split the activity and test whether any segment was found in the RMD department (e.g. "Latin America" in "Latin American Studies"). If no activity match was found, the test then looked for the RMD department name in the NAR's citation. Both tests changed all data values to lowercase to avoid any mismatches caused by differences in capitalization (see Figure 4).

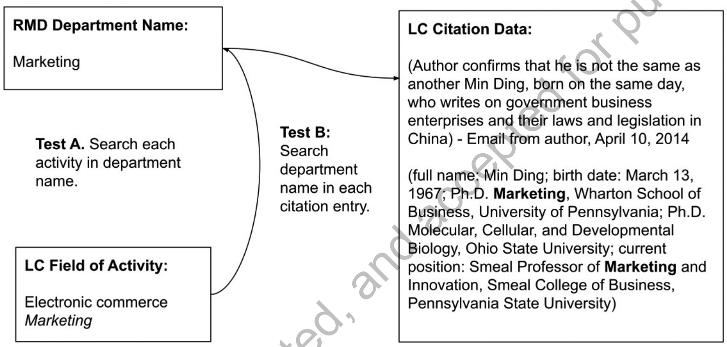
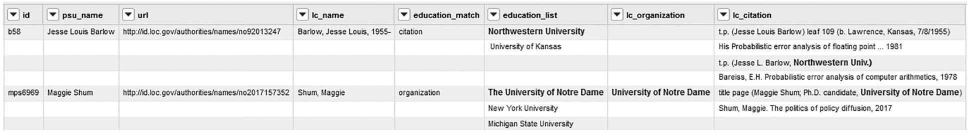

RMD Department Name:
Marketing
Test A. Search each activity in department name.
Test B: Search department name in each citation entry.
LC Field of Activity:
Electronic commerce
Marketing
LC Citation Data:
(Author confirms that he is not the same as another Min Ding, born on the same day, who writes on government business enterprises and their laws and legislation in China) - Email from author, April 10, 2014
(full name: Min Ding; birth date: March 13, 1967; Ph.D. Marketing, Wharton School of Business, University of Pennsylvania; Ph.D. Molecular, Cellular, and Developmental Biology, Ohio State University; current position: Smeal Professor of Marketing and Innovation, Smeal College of Business, Pennsylvania State University)

Figure 4. Example of testing a department with LC activities and citation data.

Output and Review

The script generated JSON records pairing LCNAF and RMD data for review. In addition to all the original data points from RMD and any LCNAF occupation, field of activity, and citation data that existed, output records contained specific fields recording whether additional match points had been found and supplemental fields from the matched NAR, including the citation. These combined JSON records were saved as a single file and then sorted by another script into batches for assessment and match confirmation. For example, one batch of records contained the highest confidence pairings with matches present for occupation, education, and Penn State affiliation.

These batches of JSON records were then imported into OpenRefine, where the record view allowed for easier review of multi-value fields (see Figure 5; additional columns were cropped for readability).



id	psu_name	url	lc_name	education_match	education_list	lc_organization	lc_citation
b58	Jesse Louis Barlow	http://id.loc.gov/authorities/names/no92013247	Barlow, Jesse Louis, 1955-	citation	**Northwestern University**		t.p. (Jesse Louis Barlow) leaf 109 (b. Lawrence, Kansas, 7/8/1955)
					University of Kansas		His Probabilistic error analysis of floating point ... 1981
							t.p. (Jesse L. Barlow, **Northwestern Univ.)**
							Bareiss, E.H. Probabilistic error analysis of computer arithmetics, 1978
mps6969	Maggie Shum	http://id.loc.gov/authorities/names/no2017157352	Shum, Maggie	organization	**The University of Notre Dame**	**University of Notre Dame**	title page (Maggie Shum; Ph.D. candidate, **University of Notre Dame**)
					New York University		Shum, Maggie. The politics of policy diffusion, 2017
					Michigan State University		

Figure 5. Example from review process showing educational institution matches.

For 2,287 of the 4,141 faculty profiles run through the testing process, 10,103 possible name matches were found, an average of 4.4 per person. However, only 1,152 of these possible matches, representing 895 possible faculty, included at least one additional match point and qualified for this project. These 1,152 matches were reviewed manually, and 649 correct matches were identified.

Assessment

Overall, the team found that limiting the possible matches using additional data points resulted in a high degree of correctness. For three types of datapoint, the presence of as little as one matched data point strongly correlated (89 percent) with a correct match. The one major outlier, as shown in Table 2, was the test for occupation.

Table 2.

Overall incidence of secondary datapoints

	Affiliation	Occupation	Education	Department
Total Found	524	700	227	296
Correct Matches	516	229	209	281
% Correct	98.5%	32.7%	92.1%	94.9%
% of all matches	79.5%	35.3%	32.2%	43.3%

Further review found that any record pairing with two matches *other than occupation* was a correct match. Additional field-by-field analysis and recommendations follow.

Institutional Affiliation

Institutional affiliation was the most widely-found field (80 percent) with the greatest likelihood (over 98 percent) of indicating a match between the faculty member and NAR. All but one of the exceptions had earned a degree, most often a doctorate, at Penn State. The other exception was a fellow member of the university's faculty, with the same name, who belonged to a different department than the person being tested.

A post-project analysis was conducted to determine how many varieties of the university's name were found in the citation field of matched NARs. These variations, represented in Figure 6, validated the authors' choice to search using a variety of abbreviations.

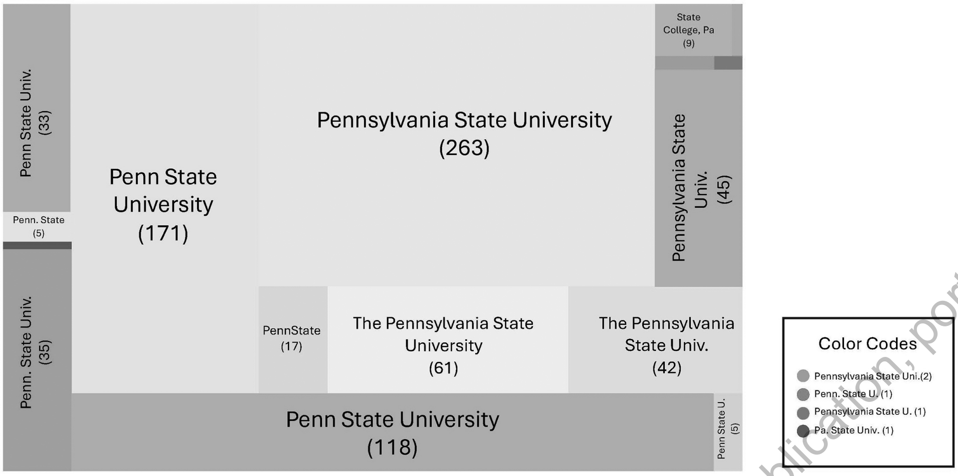


Figure 6. Distribution of institutional name variants found in citations.

The researchers recommend using the abbreviations in this article to compile an initial list of possible institutional names. Find the NARs of known faculty members and look for additional variations.

Occupation

Occupation proved, by far, to be the least reliable field in identifying matches. Of the 473 records where *only* an occupation datapoint was found, 11 were for the correct person. The challenge of common names within a large occupation pool was compounded by the flexibility of the name match, which introduced names with minor variations. The inclusion of the occupation field in the review output and the use of multiple occupation fields in NARs, for example "College teachers" and "Anthropologists," still proved useful, allowing the reviewer to filter out these mismatches quickly using local department information.

The correct match rate improved substantially when paired with a second field. Of the 199 matches that had both institutional affiliation and a matching occupation field, 197 or 99 percent were correct. The two incorrect records were both alternative institutional matches, Penn State alumni who had gone on to become faculty. Pairing the other two fields yielded slightly lower match rates. Occupation and education (95 percent) and occupation and department (97 percent) primarily found other faculty who either attended or were employed by institutions where the local faculty member was educated or worked in the same field.

The researchers note that the occupational test may still be useful if only tested when an additional match is found but recommend that it not be used in isolation. It was helpful during the review process and might be included in review outputs even if not tested.

Educational Institution

Educational affiliation was the least likely to be included but, when present, proved highly indicative of a correct match. Records that did not include affiliation with Penn

State but contained at least one of the author's educational institutions represented 83 of the 209 correct matches found. For 59 of those 83 records, it was the only datapoint present. This pool likely reflects NARs created for the author's dissertation or while they were at a prior employer.

While the team could not apply the same research and care to the programmatically generated educational institution names as to the Penn State variants above, these programmatic versions still proved useful. A review of the results found that 42 of 209 records contained only a form that varied from the faculty-entered institution name. In 10 of these cases, the variant was found in "organization," not citation, and had been caused by the faculty member prefixing the institution's name with "The" or abbreviating to "U of."

During the initial phases of the project, the team attempted to construct a variant representing possible abbreviations, as in UMD or UT. Unlike other educational matches, where comparisons were done in lowercase to avoid capitalization inconsistencies, these were capitalized. A review of these results indicated that no items matched correctly on abbreviations, and the few possible matches found were part of larger acronyms. This test variant was abandoned.

The researchers note that, if educational data is available, it may be used to find records created before the person became affiliated with their current employer. Variant forms contribute to locating matches, but abbreviations do not. The method of creating variants used for this project provides a starting point, but available data should be reviewed to determine local edge cases.

Department

The process of searching for department-related information used both the field "Field of activity" and the citation (see Table 3). Seeking the department name in the citation was much more likely to yield results, but a match in field of activity still corresponds strongly with a correct match. In 30 of 281 cases (10 percent), it was the only match point found.

One noteworthy quality of RMD department names is the lack of extraneous words (in most cases) such as "School of" or "Department of." This likely improved the number of matches found when searching the RMD department name in a NAR citation. Even without extra words, it was anticipated that complex department names, which are prone to change, abbreviation, and multiple representations, might be found less often than simple department names. However, just over half of department matches (146 of 281) found using this process were for departments whose name included two or more words. Finally, in some cases, the RMD data contained a campus name (e.g. Penn State Schuylkill) rather than a department one. While this prevented field of activity matches, a handful of citation matches were found for these records.

Of the 281 total records with correct department matches, 223 included Penn State affiliation somewhere in the record. For records that did not include an institutional affiliation, 8 matches were found in only field of activity, 39 only in citation, and 8 in both. In many of these cases, department names were represented in the citation as a description of the person's field, such as sociology. Others represented an affiliation (as a student or faculty) with a department of the same name at another institution, as in African American Studies.

Table 3.

Comparing matches found in field of activity, citation, or both

	Field of Activity Only	Citation Only	Both
Total Found	38	196	62
Correct Matches	30	190	61
% Correct	79%	97%	98%

The researchers note that both parts of this test proved useful in identifying matches. In cases when source data includes additional words such as "Department of" and "School of," it may be necessary to remove these patterns before testing.

Continuing Need for Manual Review

Despite the 100 percent correct match rate when any two of affiliation, education, or department fields were matched, the team still strongly recommends some degree of manual review. The possibilities for variation in the 95 percent match rate for names and the likelihood that people with common names (Mark Roberts or Lei Wang) might work at overlapping institutions or in the same field is low but not impossible. If a 100 percent match is desired, some review should be conducted.

An advantage of reviewing based on datapoints matched is the possibility of creating several tiers of matches. In cases where three datapoints matched, the person conducting the review may expect to spend minimal time reviewing. Possibilities with two data point matches would be similarly quick. Matches with only one matching datapoint should be expected to require more time for review. However, initial passes could be done by undergraduate student workers trained to identify and flag discrepancies for librarian review. This would require less extensive training than for reconciliation work based entirely on name matches.

Limitations

One significant limitation for comparing work done in this project with traditional OpenRefine reconciliation was that OpenRefine's match algorithm does not use a lower-bound match threshold. To test whether a different name match threshold might improve the process, the team conducted a traditional reconciliation process for a batch of 1,000 names not found during the project and identified another 360 matches. The entire set of 1,000 names were then run again through the tests outlined in this paper, with a match threshold lowered to 80. The team sought to determine which records contained identity attributes that would have led to their inclusion in the initial batch, had the threshold been lowered, and how many incorrect matches such a change would add to a review set. Results indicate that if the match threshold had been lowered to 90, the original analysis would have produced another 20 correct matches and an additional 37 incorrect matches. Lowered to a base threshold of 80, it would have produced a total of 41 additional matches but added an additional 1,172 incorrect matches.

The team identified four primary cases when a match threshold would be too low to match a person's name to their record:

- when a nickname used in the university system was not reflected as an alternative name (Jack for John),
- when surname(s) had changed over time and the current local form was not recorded as a variant on their authority record,
- when the name included a parenthetical which repeats the first name and spells out an initial ("Bell, Sarah J. (Sarah Johanna)")
- and when the university data did not contain any middle initial, the authority used the middle initial and did not include a variant without it, and the name was short enough for the ratio to be significantly lowered by this discrepancy (the pair "Li, Julie" and "Li, Julie A" have a fuzz.ratio of 86 percent).

The first two are natural limitations of working with name data. Names are represented in different ways in different contexts and may evolve over time. The third limitation could be somewhat mitigated by removing parentheticals along with dates. The fourth could be addressed with additional logic lowering the match threshold in cases where the length of a name string was below a minimum threshold. Further experimentation would be needed to determine an appropriate threshold for a given length.

In some cases, not enough data existed in either record for programmatic matching to succeed. 249 pairs from the follow-up test matched the name at 95 or higher but did not match any data from the RMD record. In some cases, this was because the NAR had been created prior to the person's Penn State affiliation and only 72 percent of RMD faculty records included educational history. NARs which pre-date the introduction of the identity attributes often only contained name and citation data, which significantly lowered the likelihood of a match. This suggests limitations in performing such matching for older sets of names or for people who had significant careers prior to commencing any of the affiliations that might be used in matching.

The reproducibility of this work will be limited by institutional infrastructure and departmental capacity. The launch of the RMD and its API made this project easy for the authors to attempt. Since RMD is not one of the "out of the box" profiling systems on the market, the authors cannot recommend how such systems might output data of similar utility. Although the scripts and processes for their use are published on GitHub, at least one member of a project team would need sufficient experience to adapt these to fit any local data.[13]

Conclusion

Libraries play an ever-evolving role in partnership with their researchers, from providing them vital access to resources, to enabling them to meet open access mandates, to showcasing their work. Data generated from pairing identifiers across library silos may be used to spotlight institutional achievement, support library assessment, and could also be of use to the institution's newer researchers, particularly graduate students, as they search for potential collaborators. This project's findings demonstrate that reconciliation efforts adding on as little as an institutional affiliation can substantially improve the quality of potential matches. Library of Congress Name Authority Records are far more than name strings; they offer a wealth of data for querying. While not every NAR or faculty profile will contain appropriate data points to support the work, the automation of these queries is possible. Reconciliation projects matching only on names should be seen as a starting point for this work and not the best that the field can do.

In cases where the local dataset only includes a name, it still may be useful to perform reconciliation using methods similar to the name matching described in this paper. Identity attributes for each possible match–all organizational affiliations, occupation, field of activity–and citation data could then be output into the data used for review. Even for the batch of occupation matches, in which a much lower number of potential matches proved correct, the team found the presence of this data saved time that would otherwise be spent clicking through to Library of Congress linked data webpages or searching for context online.

While this project focused on enriching local data, methods described herein might be adapted to support other types of authority work. For example, one might identify records which might be enhanced with better organizational affiliations by removing the query for institutional affiliation in the organization field while continuing to search for it in citation and for matches in all other fields. Catalogers could then enhance authority records with MARC 374 fields recording the person's affiliation with their current employer. One might expand outputs to include the Wikidata identifiers present in Library of Congress's Linked Data Authority File. These could be used to build a database of Wikidata records for review, enhancement, or reuse. This article provides initial methodologies and a code repository of Python Scripts which could be used to develop future work. Any such efforts will require review of local data and its suitability for testing against the external system.

Future Work

While the project described in this article demonstrates the utility of underutilized data in Library of Congress Name Authorities, the work is not easily replicable. The authors hope that the demonstration of the utility of identity attributes in this paper will inspire future work that improves data utility and lowers barriers to its use. The researchers have identified areas of systematic change that would empower others to perform similar work, places where the code may be improved, and starting points where those with less technical proficiency might engage.

As the authors were completing the first draft of this article, noted library researcher and software developer Ed Summers expressed his own frustrations with the high barrier to processing the Library of Congress's method of expressing authority records as JSON-LD.[14] The JSON-LD Framing approach he proposed and released in the idloc Python library might improve the phase of the project which required processing LCNAF's JSON-LD. The authors look forward to experimenting with the tool. However, since this project used an entire NAF download, not a specific record, the idloc library is more likely to provide a starting point for a next iteration than a solution.

It is possible that an improved serialization as proposed by Summers could make LCNAF data compatible with enhanced reconciliation, which, as described previously, cannot be used with LCNAF and many other major name authorities. The next challenge to implementing methods described here is the need to search within a field. The default method used by OpenRefine's reconciliation service is fuzzy matching. However, the reconciliation specification invites reconciliation services to develop their own methods for scoring matches. If an eventual LCNAF reconciliation source supported more than just fuzzy matching, this could significantly lower barriers to using the methods described in this article.

During the review process, the team identified two areas of potential improvement for the script. First, one might expand text matching in the LCNAF by using recursive linked data querying, acquiring textual labels at the time of processing, and searching up and down hierarchies. Such an approach might prove useful when the terms used in a person's department name and field of activity were in the same hierarchy at varying levels of specificity. However, the scope, computing power, and processing time required by such an approach made it unfeasible for this initial experiment.

Second, due to the complexity of the other challenges involved in this experiment, the project team did not experiment with emerging large language models or natural language processing libraries. As discussed in the Limitations section, the complexity of name matching posed a challenge and would benefit from ongoing revisions. These technologies might be of particular use in matching names of any length, accounting for nicknames, or addressing the problems of repeated parentheticals in authorized name forms. This would be a rich area for future research.

While this project focused on the Library of Congress Name Authority File, text string extraction and comparison of secondary characteristics demonstrated here may also prove useful when querying other data sources. For example, the authors found ORCID's API specification includes an affiliation organization parameter which may be used in queries. As with the project described in this paper, experimentation with affiliation ORCID's API also benefited from the use of variant names and parts of names. While matching one's faculty to their ORCID profiles does not solve the issue of connecting to the library catalog, it might be a more approachable starting point for someone just getting started with Python and could still enhance local data.

Acknowledgements

The authors recognize that the many catalogers who created this data, first as citations and then in structured fields, made our project possible. For those who create and update name authorities, we hope that this project demonstrates the value of your work and encourages you as you continue.

The work was supported by the Sally W. Kalin Early Career Librarianship for Technological Innovations Funds from this endowment were used to create a graduate assistantship within the Cataloging and Metadata department at Penn State's Library and to purchase a laptop with 64GB of RAM and an i9 Intel core, which was used in several projects including this one. The authors would also like to thank Anna Neatrour for graciously reading drafts of this article and providing feedback.

Ruth Kitchin Tillman is the cataloging systems and linked data strategist at Penn State University Libraries, email: rkt6@psu.edu, ORCID: 0000-0003-4547-8879.

Gala Campos Oaxaca is a doctoral candidate in educational psychology at Penn State, email: gala.campos@outlook.com, ORCID: 0000-0003-1698-3884.

Appendix A

Sample Records and Match Analysis

The following pair of sample records match on all points:

Sample Working RMD Record

```
{
    “access_id”: “kaw466”,
    “name”: “Kelly Ambrosi Wolgast”,
    “first_name”: “Kelly”,
    “last_name”: “Wolgast”,
    “inverted_name”: “Wolgast, Kelly”,
    “title”: “Assistant Dean”,
    “dept_name”: “Ross and Carol Nese College of Nursing”,
    “education_history”: [ “University of Alabama at Birmingham”, “Vanderbilt University”, “The Pennsylvania State University”]
}
```

Sample Working LCNAF Record

```
{
“id”: “/authorities/names/no2023021119”,
“authorized_name”: “Wolgast, Kelly A.”,
“activity”: [ “Nursing--Research”, “Nursing--Study and teaching”, “Education, Higher”],
“occupations”: [ “Nurses”, “Nurses”, “University and college faculty members”, “Universities and colleges--Faculty”],
“organizations”: [ “Pennsylvania State University”],
“citation_data”: [
    “faculty directory (Kelly Wolgast: Education: Doctor of Nursing Practice, December 2012, University of Alabama at Birmingham, Master of Strategic Studies, June 1995, United States Army War College, Master of Science in Nursing, May 1993, Vanderbilt University, Bachelor of Science in Nursing, May 1985, The Pennsylvania State University; Scholarly interests: military and veterans health, nursing leadership, health and wellness)”,
```

[{"@id": "https://www.nursing.psu.edu/directory/wolgast/"}, "Penn State Ross and Carol Nese College of Nursing website, viewed February 23, 2023:"],
"title page (editor, Kelly A. Wolgast) page iii (Kelly A. Wolgast, DNP, RN, FACHE, FAAN, COL (R), US Army, Associate Teaching Professor and Assistant Dean for Outreach and Professional Development, Ross and Carol Nese College of Nursing, Director, Penn StateCOVID-19 Operations Control Center, The Pennsylvania State University, University Park, Pennsylvania, USA.)",
"COVID-19 and pandemic preparedness: lessons learned and next steps, 2023:"]
}

Match Points

Name

The testing script would have initially tested Wolgast, Kelly against Wolgast, Kelly A. and found only a match of 90. It would have then tested Wolgast, Kelly Ambrosi against Wolgast, Kelly A. for an even lower match of 82. Finally, it would have tried Wolgast, Kelly A. against Wolgast, Kelly A. for a match of 100. This is above the minimum of 95.

Affiliation

Pennsylvania State University is in the organizations field and in one citation field. Penn State appears in two citation fields. This is a positive match.

Occupation

One listed occupation is University and college faculty members. This is a positive match.

Education

This is an interesting and not-uncommon case where a faculty member had received a degree from Penn State. In some cases, the presence of Penn State in their organization listing, alongside the rest of the organizations where they were educated, performed dual-duty as both affiliation and educational history. In this case, it is caught by the education test for a variant of the original data, The Pennsylvania State University, which has had its The removed. This is a positive, if unintended, match.

Citations contain University of Alabama at Birmingham, Vanderbilt University, and two separate instances of The Pennsylvania State University (one referencing education, one her employment). This is a positive match.

Department/Field of Activity

The following fields and field segments are searched in the department name: Nursing, Research, Study and teaching, and Education, Higher. Nursing is found within Ross and Carol Nese College of Nursing. This is a positive match.

The department name, Ross and Carol Nese College of Nursing, is searched within citations. It appears within one. This is a positive match.

A set of sample records, processing scripts, and instructions can be found at: https://github.com/ruthtillman/lcnaf-recon-with-attributes

Notes

1. Robert Maxwell, *Maxwell's Guide to Authority Work* (American Library Association, 2002), 1.
2. Rebecca A. Wiederhold and Gregory F. Reeve, "Authority Control Today: Principles, Practices, and Trends," *Cataloging & Classification Quarterly* 59, no. 2-3 (2021): 133, https://doi.org/10.1080/01639374.2021.1881009.
3. Dorothea Salo, "Name Authority Control in Institutional Repositories," *Cataloging & Classification Quarterly* 47, no. 3-4 (2009): 253-259, https://doi.org/10.1080/01639370902737232.
4. Karen F. Gracy, "Archival description and linked data: a preliminary study of opportunities and implementation challenges," *Archival Science*, 15 (2015): 239-294, https://doi.org/10.1007/s10502-014-9216-2; Moira Downey, "Assessing Author Identifiers: Preparing for a Linked Data Approach to Name Authority Control in an Institutional Repository Context," *Journal of Library Metadata* 19, no. 1-2 (2019): 117-136, https://doi.org/10.1080/19386389.2019.1590936; Greta Heng, Timothy W. Cole, Tang Tian, and Myung-Ja Han, "Rethinking Authority Reconciliation Process," *Cataloging & Classification Quarterly* 60, no. 1 (2022): 45-68, https://doi.org/10.1080/01639374.2021.1992554; Jeremy Myntti and Anna Neatrour, "Use Existing Data First: Reconcile Metadata before Creating New Controlled Vocabularies," *Journal of Library Metadata* 15, no. 3-4 (2015): 191-207, https://doi.org/10.1080/19386389.2015.1099989; Bria Parker and Adam Gray, "Rethinking the University of Maryland Authority File for the Linked Data Environment," *Journal of Library Metadata* 19 (2019): 69-81, https://doi.org/10.1080/19386389.2019.1589699.
5. Matthew Wright and Jennifer Carruthers, "Breaking the Bottleneck: Automating the Reconciliation of Named Entities to the Library of Congress Name Authority File," (2015) http://deepblue.lib.umich.edu/handle/2027.42/138107 ; Downey, "Assessing Author Identifiers;" Myntti and Neatrour "Use Existing Data First."
6. Myntti and Neatrour, "Use Existing Data First," 199.
7. Downey, "Assessing Author Identifiers," 126.
8. Parker and Gray, "Rethinking the University of Maryland Authority File," 75.
9. "Researcher Metadata," Github.com, https://github.com/psu-libraries/researcher-metadata/.
10. Library of Congress, "About NACO," https://www.loc.gov/aba/pcc/naco/about.html.
11. Ed Summers, "On Publishing JSON-LD," https://inkdroid.org/2024/02/14/publishing-jsonld/.
12. Penn State directory IDs are composed of 3 letters and 1-4 numbers. The letters represent the person's first, middle, and last initials at time of hire. If a person does not have or share a middle name, the letter "x" is used instead.
13. "LCNAF Recon with Attributes," Github.com, https://github.com/ruthtillman/lcnaf-recon-with-attributes.
14. Ed Summers, "On Publishing JSON-LD."