p.61→

Name Authority Control in Repositories

By

Charity K.M. Stokes, Cataloging and Metadata Librarian, Texas A&M University Libraries

David B. Lowe, Digital Collections Management Librarian, Texas A&M University Libraries

Introduction

When users search for information on a particular topic or for works by a particular person, there are certain assumptions they make. One of the most cherished is that they can easily find all works on or by a given person with minimal effort and, if they have the name wrong, that they will be pointed in the right direction as to what form of that person's name to use in their search. However, neither this assumption nor the functionality of collections comes about naturally. It is a result of how the metadata is entered and organized. One of the guiding principles of information organization is that of authority control.

What is authority control? It is a way to gather all variations of the names of a person, corporate entity, or subject into one authorized access point. Someone could be known by a nickname, their initials, a stage name, or a pseudonym, yet each variation refers to the same person.

←p.61

p.62→

Most institutional repositories (IRs) are found in academic institutions where the primary resource is the tenured faculty members whose critical career need is to be able to chronicle their work. As the push for open access gains steam, one of the ways faculty can show the impact of their work is simply by how many downloads, references, and views their work has had, which IRs can provide. But a key prerequisite is that faculty need to have their work represented under one name. If their name appears in multiple ways, it is a challenge to gather all of that information under one heading.

Pre-Repository Authority Control

One of the most interesting aspects of modern authority control is that it did not originate from user needs. It began simply as a tool, an in-house file used by catalogers so that they could bring all the works of a single person under an "authorized" name. The catalogers created the preferred heading – usually garnered from the first book by that person added to the local collection. Any other works by that author were then placed under the established heading for that library's collection.

So, in a quirk of library history, what is now a major access point started out in local technical service departments as a part of the catalogers' toolbox. Charles Cutter, the renowned creator of cataloging rules in the U.S., even stated that authority control was to be performed for the convenience of the cataloger (Cutter, 1891). In the grand old days of card catalogs, the authority file was kept in the back room of the cataloging unit. A new book would come in, the cataloger would check the author listing within the book against the local authority file and then create the bibliographic card under the heading of the author utilizing the locally authorized name.

When Cutter wrote his rules, collections were a great deal smaller, perhaps consisting of a thousand or so books. In the early 20th century, most libraries' collections included around 3,000 volumes per capita (Kevane & Sundstrom, 2014). In 2012, the average size of collection in a United States public library was 110,708 items with a median of 42,833 (Grimes, Manjarrez, Miller, Owens, & Swan, 2014). Because of the smaller size of collections, authority control was useful to the cataloger, but not obvious to the user given that few authors would share the same name as another. However, as book publication increased, so did the possibility that two or more authors would share the same name, increasing the need for authority control. However, such work is labor intensive, which limited the extent of its implementation.

In the 1930's, to address the growing need, many libraries began depending on the Library of Congress (LC) for cataloging cards. As part of this service, LC also began making cards available for the local authority files. Local name authority files began to reflect the same authorized name headings found in other libraries across the country. It did not take long before the author file cards became the accepted form of an author's name. Thus, a national name authority file was born.

←p.62

p.63→

Still, cataloging manuals at the time paid scant attention to *how* to do authority work. Cutter's Rules only say: "Give the names, both family and Christian, in the vernacular form, if any instance occurs of the use of that form in the printed publications of the author" and "when an author's name is various spelled, select the best authorized form as heading, add the variants in parentheses, and make references from them to the form adopted" (Cutter, 1891) (pp 24-25). Even the original Anglo-American Cataloging Rules (AACR), published in the 1960's, only advised to "make a heading under author's name in full and in vernacular form … enter under family name followed by forenames and dates of birth and death for specific identification when available" (Wynar, Tannenbaum, & Christensen, 1966). Neither explicitly states how to create a separate authority file, only that the cataloger is to refer to the authorized name. Catalogers developed authority files as their primary tool to know what the authorized form of names (personal and corporate) and subjects would be (Auld, 1982).

It was only with AACR2 (1978) that approved and standardized authority control practices came to be (which included an entire chapter on the *see* and *see also* references) (American Library Association, 1967). However, once again, the emphasis was on the fact that references *should* be made, but now *how* to do it. In 1978, *Authorities, a MARC Format* was published. It set a national baseline for automated authority records, based on the American National Standards Institute (ANSI) standard for the communication of authority records by means of magnetic tapes. This preliminary guidance only carried an implicit standard for quality, "with specifications and content designators for name, uniform title, and subject authorities," (Auld,

1982) (p. 323) thus establishing the differing types of electronic authority files we have today. This preliminary edition was replaced in 1981 when LC published the first edition of *Authorities,* which also included the addition of series authority and series treatment. With this publication, a national standard became available for the recording, structuring and sharing of authorities for names (personal and corporate, uniform titles, subjects and series).

Another development that was driving the push for clear standards and rules was the advent of the Online Public Access Catalog (OPAC). When Machine Readable Cataloging (MARC) appeared as a bibliographic standard in 1968, the same need for standardized names that was first felt in the infancy of modern cataloging now required even more attention as the quantities scaled upward. Users still needed to identify authors and the complete listings of their works. The only thing that changed was the scope of the bibliographic universe. No longer was it just about what the local collection contained, but the entire library community.

←p.63

p.64→

It was not just libraries who were concerned about authority control; the publishing industry was no less affected. Bowker was one of the very first to compile an authorized list of authors. With their iconic publication *Books in Print*, Bowker also needed to know that the books written by one author were attributed to that author. So, in 1981 Bowker published *Authors' Names; An Authoritative Listing of Personal and Corporate Names,* which was based on LC records (Bowker, 1981).

It was the development of the computer age that accelerated these developments. Catalogers began sharing their knowledge as well as their records via cataloging utilities. The Library of Congress Authority File (LCNAF) became the definitive 'authority file' for the country and most of the western hemisphere. By using a standardized and trusted source, libraries reduced the overhead for cataloging by doing away with local authority files. It increased productivity and reduced the cost associated with cataloging.

Authority control changed tremendously with the introduction of *Functional Requirements for Bibliographic Records (FRBR)* (International Federation of Library Associations, Study Group on the Functional Requirements for Bibliographic Records., 1998) and the *Functional Requirements for Authority Data (FRAD)* (IFLA Working group on Functional Requirements and Number of Authority Records (FRANAR), 2008). This was a radical shift away from Cutter's requirements to identify and disambiguate objects of a catalog to fulfilling the specific FRBR user's tasks. The tasks are

- To **find** entities that correspond to the user's stated search criteria
- To **identify** an entity
- To **select** an entity that is appropriate to the user's needs
- To acquire or **obtain** access to the entity described

Within Repositories

To relate this historical narrative with IRs, consider that, with digital collections, the operative word is ‘collections’. A library’s collections need to be accessible. Just as libraries provided access through book indexes, card catalogs, OPACs, etc., metadata librarians create a surrogate record (metadata) in order for users to locate the information they seek. One of the most important aspects of what users need – especially academic users – is to find all the works by an author, or any works that a person has contributed to. To this end, the metadata for authors (and subjects and series) needs to be collocated – gathered under a consistent, authorized form. The tools we use may have changed, but the needs of our users have not.

←p.64

p.65→

Focusing on the IR context, a worthwhile preliminary discussion might include the evolution of the terms “digital collections” and “institutional repository” and the library functions associated with them—such as they have come to be—since the advent of the Web. Initially, digital collections as a phenomenon in academic libraries grew largely out of in-house scanning operations that pre-date networked information tools like web browsers. The content sources were holdings from the libraries’ archival and circulating collections: photographs, postcards, maps, and most certainly print items such as books and journals, but also audiovisual materials like local oral histories. Such digital conversion activities had begun in libraries before networks blossomed in the 1990s, with file sharing happening via the various evolving media of the time. After the release of the first graphical tool Mosaic in 1993, browsers became the obvious frame for all of this content. At the same time, capture equipment to produce that digital content from analog objects (like books and paper) became more affordable, which led to more content getting created and available. The following decade saw the rise of mass digitization efforts, particularly of textual objects, such as those led by the Internet Archive, Microsoft, and the Google Books project, parts of all of which have been collocated spectacularly in HathiTrust.

Concurrently with this conversion from paper, responsible information professionals also devoted their attention to born digital materials. Purists would argue that a true IR consists primarily of “born digital” materials, predominately electronic theses and dissertations (ETDs), articles, conference papers, and presentations; “digital collections,” on the other hand, they use to refer more to sets of items digitized from analog originals, perhaps mixed with some born digital materials. The first university to require electronic submission of theses and dissertations was Virginia Tech, in 1997 (https://vtechworks.lib.vt.edu/handle/10919/5534), but the true arrival of IRs may be comfortably dated to the initial development and release of the DSpace software in 2002-2003, mentioned by Clifford Lynch in his seminal work advocating for IRs. (Lynch, 2003)

One feature of IRs that complicates the metadata aspect is that much of the ingested content (OA articles, gray literature) was not as routine to cataloging workflows and the staff that ran them. If they were even involved at all with the new IR materials workflows, typical cataloging departments at the latest turn of the century were heavily involved in monograph, serial, and A/V materials cataloging, but were less accustomed to material without a book-like title page or “chief source of information,”—as the trade terminology goes--which typically would have fallen to more specialized and experienced “original catalogers,” as they are known. The sheer numbers of these new documents, the lack of formal identification of their authors, and unmediated deposit (with authors or their designates left to their own devices), has led to an accumulation of content in IRs that seems utterly devoid of traditional authority control.

←p.65

p.66→

It was considered as an advantage that the authors could input their own metadata. Structurally, the idea was still the same: describe the work and provide access points. And who could better describe a work than the authors themselves? Having authors or their designates essentially catalog their own works would cut out the middle man and allow faster access to material.

Such a tack was functional for many years. Basic access fields such as author and title were present. However, because there was no authority file, variations in names began to crop up. At first, the variations were a minor inconvenience. But just as with physical collections, as digital collections grew, so did the tangles created by the lack of authority control. Users began to grow frustrated trying to figure which of the five Professors J. Vance were they looking for. Was it James Vance? Or Joan Vance? Jack, Jill, or John? And if it was a James Vance, did that include James A. Vance and was he the same as James Allen Vance? Which one was the composer and which was a respected professor of biology? Users became frustrated and would often would give up when trying to find the works they were looking for by a given author. There were variations due to the fact that occasionally authors would use their full names, while at other times their initials. And then there is the issue simply of having an extra space or a misspelling, creating endless variants.

Challenges

A suitable review of the topical literature would be a summary of the challenges reported in the scholarly record related to authority control in IRs. The relatively brief history of IRs, in tandem with the concentration of early related work being heavily devoted to advocacy for IRs as a concept, translates into a quick exercise. Among the first to call attention to the issue of authority control in IRs in a comprehensive, studied manner was Salo (Salo, 2009) who outlines problems with the available tools and related workflows but also speculates about some fruitful paths toward resolutions.

One issue brought up by Billey (Billey, 2019) has to do with privacy. Current standards for the industry standard Name Authority Cooperative Program (NACO) authority records are based on FRAD, which greatly expanded the number of attributes to describe people. Under Cutter’s rules, the only need for attributes was to identify and disambiguate the names in order to facilitate discovery. However, when FRAD was codified into Resource Description and Access (RDA) in 2010 and established as the standard through the RDA Toolkit (Joint Steering Committee for Development of RDA., Chartered Institute of Library and Information Professionals., American Library Association., & Canadian Library Association., 2010), catalogers

←p.66

p.67→

went from disambiguation to actually describing people. This descriptive process included not just information that was already in use (name of person, dates such as year of birth and/or death, fuller form of name) but to much more detailed information such as: gender, country, address, profession, titles, and affiliations.

On the surface, many US based academic institutions would see no problem with including the above information. However, given the current political atmosphere, something seemingly innocuous as affiliation can cause personal problems. For example, perhaps there is a popular children's author who is member of a minority religion. This personal affiliation has nothing to do with their writings or works, but can result in them being targeted by the more radical elements that exist in society. Then there is the issue of gender. If the person's gender – especially if non-binary – impacts their work, it will be obvious in the work itself. If not, it really does not serve any purpose in the identification or disambiguation of the author's name. So the challenge for authority control in repositories is the same as for many online media – that of balancing the privacy of the individual against the needs of the organization.

A final challenge faced by repositories is pragmatic. Budgetary constraints affect the ability to manage metadata by limiting the number of resources available to perform that management. It has long been known that authority control by people is an expensive investment in bibliographic control. Shrinking technical service departments and the outsourcing of those tasks has led to a dearth of expertise and a shrinking number of people able to perform the work. However, one possible solution is a type of human-machine hybrid system, leveraging software to help control costs (Liu & Qin, 2014).

Current State of Name Control

The authors' knowledge of practices springs from their experiences at a variety of institutions, combined with a focused search for IR workflow instructions across many institutions in 2019. It confirms that metadata for ETDs are typically handled manually by staff. When it comes to the names of authors, advisors, and committee members, the staff normalize names by referencing standard authorized forms (*e.g.*, LCNAF). If a name does not yet have an authority record, the staff usually follow a standard algorithm, such as "LastName, FirstName MiddleInitial." Sadly, even the best of algorithms can lead to conflicts, due to ambiguities introduced with features such as compound last names or life events (such as marriage) that result in name changes. ←p.67
p.68→ Less mediated workflows, such as self-deposited articles, manuscripts, or presentations are even more prone to conflicting entries. External tools such as OpenRefine (Carlson & Seely, 2017) may be enlisted to help resolve existing entries, while some IR software has limited internal functionality in this arena. For example, EPrints (EPrints Project, n.d.); (Salo, 2009) and DigitalCommons (Edwards, 2018) include the ability to merge name records.

Some authority control solutions are arising at the institutional and regional consortium levels.

Digital Library efforts at the University of North Texas (UNT) have produced the UNT Names App (https://digital2.library.unt.edu/name/). Covering primarily personal and organizational names, the application is incorporated into their IR workflows, which feature mediated deposit. Similarly, librarians at the University of Houston (UH) Libraries have implemented an instance of iQvoc called Cedar that covers the names of individuals and organizations, as well as subject terms. At Columbia University Libraries, one notable function of their metadata editing tool Hyacinth is its capacity to mint URIs for named entities that lack them elsewhere. A consortial project at the Mountain West Digital Library (MWDL), known as the Western Name Authority File (WNAF), seeks to create a central file for its partners that will be compatible with Linked Open Data (LOD) efforts.

Future Trends: LOD

Although LOD has not been mentioned heretofore in this chapter, it did not just pop up by happenstance. While having a name authority control system in place is a worthy, practical cause in and of itself, in fact there are much larger implications. Name authority systems that produce and manage uniform resource identifiers (URIs) can relate their efforts to LOD developments; in turn, the LOD efforts are connected to Artificial Intelligence (AI) and Machine Learning (ML)—some of the most promising and awe-inspiring accomplishments of the current age, as components of autonomous vehicles and smart speakers, just to name some examples. At the heart of structuring these advances for further success are open standards such as Resource Description Framework (RDF) which relies on a basic tripartite grammar structure "Subject/Predicate/Object" for which the subjects (or actors) and the objects (or those acted upon) are most commonly represented by URIs. No less than Sir Tim Berners-Lee (2006), in his seminal and foundational piece on Linked Data, states expectation #1 as: "Use URIs as names for things." So the austere world of authority control from librarianship brings us face to face with the future of humankind's interaction with information technology in our everyday lives.

←p.68

p.69→

To be coldly realistic, though, such a future is fraught with huge risks where privacy and personal security are concerned. Insofar as librarians can collectively influence the directions at hand, it is worth considering what is at stake. As mentioned previously, Billey (2019) shares sober advice about the vulnerable zone where authority work could impinge upon Personally Identifiable Information (PII). Referring to particular date, gender, and affiliation data points, Billey cautions: "Recording this information could violate a person's privacy, make their personal information vulnerable to bad actors, and even possibly put someone in danger" (pp. 10-11). The solution offered is to be circumspect in what data points even get scoped, much less recorded, in authority control systems.

Having considered the promises and risks for authority control systems related to local IR implementations, it bears emphasizing that local systems will tend to be the most relevant and familiar with the institutional context. Open URIs that unfold thorough, accurate, yet not overly revealing pockets of information about people and their groupings will have the capacity to integrate IRs with the greater world of scholarly communication and facilitate positive interaction. Standardization will be critical, but at the moment those standards are not yet established above any threshold of collective refinement. The near term ahead will be full of experimentation and surely some solid best practices will emerge.

Conclusion

For the user, the need to be able to identify all the works by a given author and know that this author is the correct one has not changed since Cutter's day. What has changed is the sheer number of authors available in the bibliographic universe. For institutional repositories, this includes not just the traditional book author, but also authors of articles and grey literature. The scale is overwhelming for the metadata specialist. For a user, it can easily go beyond overwhelming to baffling. This chapter covered the development of authority control as well as the constraints and challenges inherent in trying to impose authority control.

Authority control is a needed tool for our repositories and digital collections. Given the emphasis in academia on citations and analytics for the purposes of career advancement, faculty need to cite metrics connected to how often their works were taken up in their professional communities in order to convey impact. Graduate students would like to know what disciplines a professor has published in, as well as which professor served as advisor for other graduate students' works. No one wants to wade through multiple publications for one person who appears with as many as 12 variant names! So much needed information is lost in such a welter of name variants.

←p.69

p.70→

The rapidly developing semantic web enables a world where users can gather the information they need by following the connections between different entities – whether a person, a group, or a subject. The vision of linked data is the foundation upon which current standards (RDA, FRBR, FRAD) have been built to work in a world of artificial intelligence and machine learning.

Our world is in the middle of a revolution, an information revolution that is no less a seismic shift than the industrial or technological revolutions before it. Digital collections and repositories are not only a product of that revolution but are helping to drive it. Authority control will be part of the steering.

References

American Library Association. (1967). *Anglo-American cataloging rules: North American Text.* (C. S. Spalding, Ed.) Chicago, IL: American Library Association.

Auld, L. (1982). Authority Control: an eighty-year review. *Library Resources & Technical Services*, 319-330.

Berners-Lee, Tim. (2006). Linked Data. *Design Issues.* Retrieved from: https://www.w3.org/DesignIssues/LinkedData.html

Billey, A. (2019). Just because we can, doesn't mean we should: An argument for simplicity and data privacy with name authority work in the linked data environment. *Journal of Library Metadata*. doi:10.1080/19386389.2019.1589684

Bowker. (1981). *Authors' names: An authoritative listing of personal and corporate names.* New York: Bowker.

Carlson, S., & Seely, A. (2017). Using OpenRefine's reconciliation to validate local authority headings. *Cataloging and Classification Quarterly, 55*(1), 1-11. Retrieved from https://doi.org/10.1080/01639374.2016.1245693

Cutter, C. A. (1891). *Rules for a dictionary catalogue.* Washington: Government Printing Office.

Edwards, L. (2018, June). Authority control in digital commons: Why bother? *Presented at the Digital Commons Southeatern User Group Meeting*. Johnson City, Tennessee. Retrieved from https://encompass.eku.edu/fs_research/280

EPrints Project (n.d.). *Authority lists - EPrints Documentation*. Retrieved from https://wiki.eprints.org/w/Authority_Lists

Grimes, J., Manjarrez, C., Miller, K., Owens, T., & Swan, D. W. (2014). *Public Libraries in the United States survey: Fiscal Year 2012.* Washington: Institute of Museum and Library Services.

←p.70

p.71→

IFLA Working group on Functional Requirements and Number of Authority Records (FRANAR). (2008). *Funcational Requirements for Authority Data: A conceptual model.* München: KG Saur. Retrieved from https://www.ifla.org/files/assets/cataloguing/frad/frad_2013.pdf

International Federation of Library Associations, Study Group on the Functional Requirements for Bibliographic Records. (1998). *Funcational requirements for bibliographic records: Final report.* Munchen: K.G. Saur. Retrieved from https://www.ifla.org/files/assets/cataloguing/frbr/frbr.pdf

Joint Steering Committee for Development of RDA., Chartered Institute of Library and Information Professionals., American Library Association., & Canadian Library Association. (2010). *RDA Toolkit.* [Chicago}, IL: American Library Association.

Kevane, M., & Sundstrom, W. A. (2014). The development of public libraries in the United States, 1870-7930: A quantitative assessment. *Information & Culture, 49*(2), 1117-144.

Liu, Z., & Qin, J. (2014). An interactive metadata model for structural, descriptive, and referential representation of scholarly output. *Journal of the Association for Information Science and Technology*, 964-983.

Lynch, Clifford A. (2003). Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. *ARL: A bimonthly report on research library issues and actions from ARL, CNI, and SPARC.* Retrieved from: https://www.cni.org/wp-content/uploads/2003/02/arl-br-226-Lynch-IRs-2003.pdf.

Salo, D. (2009). Name authority control in institutional repositories. *Cataloging & Classification Quarterly*, 249-261. doi:10.1080/0613937092737232

Wynar, B. S., Tannenbaum, E., & Christensen, C. (1966). *Introduction to cataloging and classification: A teaching guide with illustrations of major principles for descrptive cataloging and classification* (Second, revised and enlarged ed.). Denver, CO: Colorado Bibliographic Institute.

←p.71