CHAPTER 19

LEARNING FROM EACH OTHER: RECIPROCITY IN DESCRIPTION BETWEEN WIKIPEDIANS AND LIBRARIANS

Angela Yon[1] and Eric Willey[1]

[1] Illinois State University

Abstract

Librarians, archivists, and museum professionals are increasingly realizing the value of using and contributing information to Wikipedia through projects such as edit-a-thons and the 1Lib1Ref project. As the amount of knowledge in Wikipedia and Wikidata grows, the benefits to libraries in partnering with Wikimedia projects to enhance their own bibliographic records and catalog search results also increase. Conversely, librarians have created an immense number of bibliographic and authority records that Wikipedia and Wikidata editors can use both as resources in and of themselves and as examples of various approaches to metadata and knowledge creation. Despite some challenges there are numerous benefits for working to integrate library data with Wikipedia more closely.

This chapter will serve to highlight differences between Wikipedia resources and library catalog records, and how librarians and Wikipedians can learn from each other to improve description and discoverability in both Wikipedia and library catalogs for their respective users. It will also illustrate differences between

these two systems in order to reduce confusion and errors when data are merged uncritically. The discussion draws on experience gained from a previous Illinois State University Research Grant-funded project that used the Wikipedia List of African-American writers to enhance library catalog records.

Keywords

Authority control, Library of Congress Demographic Group Terms (LCDGT), Wikipedia lists, Metadata, Data integration, Cataloging, Wikidata.

Introduction

Librarians, archivists, and museum professionals are increasingly real-izing the value of using and contributing information to Wikimedia projects, and as the amount of knowledge in Wikipedia and Wikidata grows, the benefits to libraries in partnering with Wikimedia projects to enhance their own bibliographic records and catalog search results increase. Librarians, archivists, and museum professionals have also created an immense number of bibliographic and authority records that Wikipedia and Wikidata editors can consult as information resources and examples of how to organize knowledge. Differences between Wikipedia resources and library catalog records provide opportunities for librarians and Wikipedians to learn from each other and improve description and discoverability in both resources for their respective users. The following discussion describes experiences gained from a previous Illinois State University Research Grant-funded project that explored using the Wikipedia List of African-American writers (Wiki-pedia contributors, 2020a) to enhance MAchine-Readable Cataloging (MARC) records with demographic group terms for authors.

Trends for the library catalog currently integrate the discoverability of local resources with features of the larger web environment. This mixture often draws from existing metadata in library catalog records. Examples include allowing users to refine searches using facets, using Functional Requirements for Bibliographic Records (FRBR) by show-ing a work in its representation of versions and editions, using linked data approaches for common entities, and integrating community-

created systems like Wikipedia (Dempsey, 2012). As the largest library cooperative, OCLC has undertaken several collaborative partnerships between Wikipedia and libraries, such as the Wikipedia Visiting Scholar program and Project Passage (OCLC Research, 2020). OCLC has also urged catalogers to "integrate researchers' external IDs within library applications and services as appropriate" to facilitate the creation of high-quality linked data between resources (Smith-Yoshimura et al., 2014).

In recent years libraries have undertaken attempts to integrate the library's catalog data into the larger web environment for discoverability purposes. An additional goal for libraries is to share and benefit from knowledge created by larger community-based open systems, platforms, and hubs such as Google Search, Wikipedia, Amazon, LibraryThing, and Google Books, by bringing them into the library catalog setting (Dempsey, 2012). The open-source library catalog, VuFind, offers optional features that allow users to view rich linked data content, such as author biographies via Wikipedia (VuFind 4.1 Milner Library, 2020). Similarly, to improve the quality of services for both libraries and Wikipedia, Joorabchi and Mahdi (2018) designed a software system for automatic mapping of FAST subject headings that are used to index library materials to their corresponding articles in Wikipedia. Charting connections between the library catalog and other open systems, such as Wikipedia, creates a need for the implementation of linked data elements. The merging of data from different systems and its many descriptive forms under one discovery layer calls for linked data approaches so that the resources may be discoverable based on common entities and identifiers (Dempsey, 2012).

Both libraries and Wikipedia generate projects that allow users to refine searches with facets, lists, and categories. In 2013, the Library of Congress began exploring the creation of the Library of Congress Demographic Group Terms (LCDGT)-controlled vocabulary (Library of Congress, Policy and Standards Division, 2020). Through inclusion of new MARC fields in bibliographic records, the terms would allow catalogers to describe intended audiences and the creators of works. Library of Congress Subject Headings (LCSH) and their subdivisions

already included information describing audiences and creators of resources (including demographic groups), but the format of the strings was not always clear to users in search results. With the use of LCDGT, there could be more precision in search results by faceted displays using these terms in the catalog, and clarity in the descriptions of the resources for users. Similarly, Wikipedia contains many lists of individuals in various demographic groups, often associated with a profession. Many of the Wikipedia lists correspond with the nine categories of the LCDGT vocabulary, one of which is ethnicity/culture, which may indicate an agreement on what categories are useful between the two systems.

From 2017 to 2019, the authors led a project to examine the degree of agreement between the Wikipedia List of African-American writers (Wikipedia contributors, 2020a) and Library of Congress criteria for determining if a creator would be considered appropriate for description using the LCDGT term African Americans. For the project, African American history subject expert Trumaine Mitchell found that there was a high level of agreement between individuals on the Wikipedia list and those whose resources might be described as being authored by an African American by the LCDGT criteria (Willey and Yon, 2019). From that project, additional lessons were learned about differences in the structure of information between Wikipedia (especially Wikipedia lists) and traditional (MARC) library cataloging.

At the time, the principal investigators were researching the degree of agreement between Library of Congress criteria and decisions by Wikipedia editors as to which writers could be considered members of the demographic group African American. The possibility of using Wikidata or Wikipedia lists updated by bots such as Listeria to populate catalog search results was not considered during this research in favor of determining if the LCDGT criteria led to the same conclusions as those reached by Wikipedia editors. If there had been disagreement, that would have been a warning flag against integrating the two platforms; however, thankfully there was not. As this article reflects lessons learned during this project, there is limited discussion of Wikidata, although it represents a wealth of possibilities for additional research.

Similarities and Differences between
Wikipedia Lists and MARC Cataloging

The initial barriers for participating in the Program for Cooperative Cataloging (PCC) versus Wikipedia differ considerably. The PCC requires institutional participants to undergo training through the PCC Secretariat before creating or editing Name Authority Records (NARs). However, creating Wikipedia lists only requires familiarity with word processing software, so most people will be able to use the visual editor to create and make edits to Wikipedia lists and pages with minimal or no additional training, although several tutorials and guides are provided for users. In the analysis of Wikipedia lists, the subject expert was quickly able to learn how to add the Authority Template and the Library of Congress Control Number (LCCN) to Wikipedia pages where it was lacking and did so with accuracy and efficiency; however, no attempt was made to train them on the creation of NARs or the editing of existing NARs because of the greater amount of time required to learn Resource Description and Access (RDA) standards, International Standard Bibliographic Description (ISBD) punctuation, and other cataloging skills.
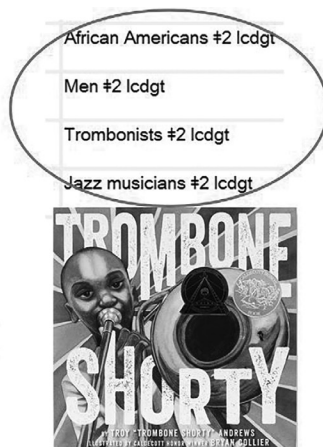
The process for making changes to PCC cataloging policy and Wikipedia policy also differs significantly. Partway through the creation of NARs in the project, PCC announced a moratorium on the use of the MARC 024 Other Standard Identifier field (Frank, 2018). In November 2020, PCC ended the moratorium and provided guidelines on the use of the MARC 024 field to link NARs to Wikidata identifiers, two years after the project. This allows NARs to link directly to Wikidata items, which are also used by the Authority Control Template in Wikipedia articles to provide links to NARs and other identifiers such as the International Standard Name Identifier (ISNI) and Virtual International Authority File (VIAF) (Wikipedia contributors, 2020b). No impactful changes to Wikipedia policy were encountered during the project, but it is understood that proposals can be made and implemented relatively swiftly if approved by the community. This is not intended as a critique of the PCC deliberative process but may be seen

as an indicator that implementing changes in older established library standards such as MARC, which has undergone several changes and updates since it was developed in the 1960s, may require more deliberation and testing than changes to a relatively new system such as Wikipedia (developed in 2001). It may also be an indication that this is a larger conceptual step for cataloging systems than it is for Wikipedia.

A similarity between the two systems is that both Library of Congress and Wikipedia require citation of evidence to show why a person is described using an ethnic or racial group in some instances but not in others. In Wikipedia, the List of African-American writers includes a note to consult the Who is African American section (which has undergone several renamings since a section by that exact name was last present in 2012) of the African Americans article (Wikipedia contributors, 2020c) and the individual pages should include citations to reliable resources justifying any claims of race or ethnicity. This can, however, lead to cases such as Stanley Bennett Clay (Wikipedia contributors, 2020d) where they are included in the List of African-American writers (Wikipedia contributors, 2020a), but their page does not describe them as Black, African American, or any equivalent term, and they are not listed in the category: African-American writers page (Wikipedia contributors, 2020e). PCC policy also requires that NARs include a MARC 670 Source Data Found field for demographic information included in the record at the time of creation; however, catalogers can edit bibliographic records and include demographic information in the MARC 386 Creator/Contributor Characteristics field without the requirement to include citations showing how they reached that decision, as shown in figure 1. Therefore, both institutions can be said to have requirements that users cite information supporting any addition of ethnic group information to certain records, and practices that specifically associate individuals with an ethnic group, but do not require citations to convey that information.

Additionally, there are differences in the structure of the LCDGT and Wikipedia lists. The LCDGT are generally broken down to a single facet, because they are intended to be used in individually repeatable

Figure 1 This image depicts Trombone Shorty by Troy Andrews, a bibliographic record, OCLC number 880349715 (OCLC Connexion, 2020).

MARC fields. Typically, two separate traits would be described using two different terms, one for ethnicity or culture and a second for nationality. The term "African Americans" is in the ethnic/cultural category but also describes a nationality. Therefore, the term "Americans" will also be included in a record. There is also no LCDGT for the occupation writer or author because creation of a bibliographic record is based on literary warrant. The overwhelming majority of entries in the Name Authority File (NAF) and bibliographic records are by writers, making that criteria nearly useless for sorting. Wikipedia lists cover many topics, but lists of people often seem to combine the criteria of nationality and profession (Puerto Rican comedians, for example). In order for library catalogs to incorporate Wikipedia lists into search results, these different approaches will need to be reconciled. Depending on how difficult this is, libraries may instead choose to incorporate information from Wikidata items or lists generated and automatically updated by tools such as Listeria (Manske, 2015). It may be easier for Wikipedia to generate lists from library records, as they can combine the individual facets to form a list with as many characteristics as desired. This may also indicate a difference in design philosophy with librarians expecting users to utilize facets to narrow search results

in a library catalog, and Wikipedia users creating lists and categories with the expectation that users will engage in something more akin to browsing through search results.

Users of both systems face difficulty in creating complete and comprehensive lists of members of specific demographic groups. It was discovered that two authors with works in the local library catalog were not on the List of African-American writers but have Wikipedia pages. Benjamin Griffith Brawley was a prominent African American author and educator, and several of his books were standard college texts in the early twentieth century. Phillip Hayes Dean, an African American playwright, also has a Wikipedia page but was not on the List of African-American writers. Both have bibliographic records in the local library catalog with LCSH terms that included African American authors. This suggests that library catalogs may be useful in either populating or at least providing initial leads on populating demographic group lists, although they will only reflect members of that group for whom the library has holdings. Wikipedia also includes both Wikipedia lists and Wikipedia categories (and there are Wikidata items as well), and users may not always update all of these leading to lists and categories describing the same group but which include different items.

The Wikipedia List of African-American writers also included authors whose library bibliographic records did not record their status as African Americans, of course. One such author provided an interesting example in how Wikipedia lists can be useful in discovering works not directly cataloged by librarians. Clarissa Minnie Thompson Allen was included on the List of African-American writers but did not have an NAR or catalog record for their novel, *Treading the Winepress*. Investigation revealed that Allen's novel was serialized in *The Boston Advocate*, a newspaper, and not published as a stand-alone work. The portion that could be located has since been printed as an open-access book by the Illinois State University publications unit (Allen, 2019), but as catalogers rarely create records for individual parts of newspapers, Allen's work was never cataloged on its own record, and no NAR for Allen was created until long after the publication of her work. Historically marginalized creators often turned to publishing their works

in formats other than monographs. Wikipedia lists can be useful in locating creators and works that may not have individual bibliographic records in the library catalog.

Mapping data to fields in a bibliographic record from similar categories in each system is an obvious example of how libraries and Wikipedia can provide each other with additional information to draw from; however, Joseph (2019) suggests a fresh approach of how Wikipedia can contribute to the library catalog. One of the challenges library catalogs face is the loss of historical revisions to bibliographic records. This change came consequently from the physical card catalog transition to the digital library catalog. "Analyses that were possible with physical catalog cards can no longer be performed, and tools that process digital records leave no traces of the information they add, remove, or update" (Joseph, 2019). In 2015, OCLC stopped printing catalog cards. Revisions and the historical context of classification are omitted in the online catalog, removing a source that librarians could reference for past analysis. Wikipedia, on the other hand, allows users to track changes in its digital environment through its discussion pages and revision pages. Joseph (2019) believes the library catalog can benefit from a similar practice, allowing analysis of changes and a larger field of subject domain experts to contribute to metadata decisions through discussion.

Library employees have easier access to databases, reference works, and special collections or archival materials than some Wikipedia members, which prove especially valuable in satisfying notability requirements for articles. While the Internet removes many barriers to access, older print materials are still largely held in libraries. Similarly, special collections materials are often only available through intermediaries or by on-site visits. Libraries also feel incentivized to provide citations from their special and local collections to bring greater visibility to those materials. In the analysis of Wikipedia lists, the subject expert began their search in Google Books but also utilized interviews found in library databases to conduct their research.

The volunteer nature of Wikipedia also makes it an excellent source of editors with rich and varied subject knowledge. Domain experts from

around the world can apply their extensive knowledge to articles and lists, at their own discretion and convenience. Catalogers are also subject experts but will likely be expected to work on materials purchased by other librarians. Wikipedians' volunteer status allows them relative freedom in choosing topics to contribute. While librarians generally must justify metadata created during their work time to stakeholders, Wikipedians can investigate topics and create lists on subjects of their choice. For the analysis of Wikipedia lists, reference librarians stated that patrons sometimes requested works by African American creators, which gave the project more credibility when composing the grant request. It is also unlikely that the project would have proceeded beyond the theoretical phase without grant funds to hire a subject expert in African American history.

Wikidata and Future Work

There is consensus among institutions that the future of this reciprocal relationship with data will be very advantageous and valuable as the catalog moves to new forms of discovery in libraries (Bartholmei et al., 2016). In 2019, the Association of Research Libraries released a white paper by a task force of library professionals and expert Wikidata users with recommendations for librarians to use Wikidata to advance discovery of their collections, faculty, and institutions. Many cataloging systems do not produce linked data and cannot make data available as open linked data. Research libraries may lower this barrier with participation in the Wikidata community and infrastructure (Association of Research Libraries, 2019).

While the project under discussion used a list from Wikipedia, Wikidata offers a low-barrier, high-result method for creating and using linked data in libraries. It makes data not only visible but also reusable as linked data. In a 2016 International Federation of Library Associations and Institutions (IFLA) discussion paper, Stephan Bartholomei and others noted "the potential of Wikidata to draw linked open data and linked open data authorities together across the world's languages and many different ontologies and taxonomies has enormous potential to support researchers around the world" (Bartholmei et al., 2016).

The Library of Congress (LC), recognizing the potential of Wikidata as being a hub of identifiers, included links in their authority records out to Wikidata in spring 2019. They bulk loaded 400,000 more LC identifiers into Wikidata to add to the 650,000 IDs in Wikidata. This brought their total to about a million of their identifiers in the system. The majority of these identifiers are to their NAFs and 35,000 link to the Library of Congress subject heading file. Likewise, these links to Wikidata also appear on over one million Library of Congress Linked Data Service authority pages and in the data (Ferriter, 2019).

The PCC also acknowledges how Wikidata can be an important collaborative partner and system to help in the development of identity management and identifier creation for libraries and institutions. In September 2020, the PCC launched a Wikidata pilot project "to further advance the movement toward identity management" (PCC, 2020). Over seventy academic and cultural institutions across the globe will be part of the pilot to increase the movement toward identity management, and membership in PCC is not required to participate in the project.

Conclusion

Even though the project had a narrow scope (focusing on one Wikipedia list and MARC cataloging), the authors were able to learn many significant lessons about Wikipedia practices, cataloging, and how they interact. The practices and goals of catalogers and Wikipedians are often aligned, and even differences between the two group's practices can be seen as complementary rather than opposed. The Wikipedia-focused project also provided an excellent entry for the authors into associated services such as Wikidata and has led to further projects using that platform. With major institutions such as PCC backing Wikidata-related projects and Wikipedians-in-residence becoming increasingly accepted, additional opportunities for collaboration between Wikipedia and academic libraries are emerging. Critically, the reciprocity of knowledge and expertise between librarians and Wikimedians can significantly improve services and contribute greatly to the overall information landscape.

References

Allen, C. (2019). Treading the Winepress; or, a Mountain of Misfortune. *Undiscovered Americas*. https://web.archive.org/web/20200927134819/https://ir.library.illinoisstate.edu/ua/2/.

Association of Research Libraries (ARL) Task Force on Wikimedia and Linked Open Data. (2019). "ARL white paper on Wikidata: Opportunities and recommendations." *Association of Research Libraries*. April 18, 2019. https://web.archive.org/web/20201110232835/www.arl.org/wp-content/uploads/2019/04/2019.04.18-ARL-white-paper-on-Wikidata.pdf.

Bartholmei, S., Franks, R., Heilman, J., Joseph, M., McDonald, V., Raunik, A., Ridge, M., & Robertson, M. (2016). Opportunities for academic and research libraries and Wikipedia. https://web.archive.org/web/20200522082117/www.ifla.org/files/assets/hq/topics/info-society/iflawikipediaopportuniesforacademicandresearchlibraries.pdf.

Dempsey, L. (2012, December 10). Thirteen ways of looking at libraries, discovery, and the catalog: Scale, workflow, attention. *EDUCAUSE*. https://web.archive.org/web/20201106232835/https://er.educause.edu/articles/2012/12/thirteen-ways-of-looking-at-libraries-discovery-and-the-catalog-scale-workflow-attention.

Ferriter, M. (2019, May 22). Integrating Wikidata at the Library of Congress. *The Signal*. https://web.archive.org/web/20201121143752/https://blogs.loc.gov/thesignal/2019/05/integrating-wikidata-at-the-library-of-congress/.

Frank, P. (2018). "024 (Other Standard Identifier) data in NACO records: Temporary moratorium." *PCCLIST@LISTSERV.LOC.GOV*, September 13, 2018. https://web.archive.org/web/20201202161342/https://listserv.loc.gov/cgi-bin/wa?A2=ind1809&L=PCCLIST&P=38986.

Joorabchi, A., & Mahdi, A. E. (2018), "Improving the visibility of library resources via mapping library subject headings to Wikipedia articles." *Library Hi Tech*, *36*(1), 57–74. https://doi.org/10.1108/LHT-04-2017-0066.

Joseph, K. (2019). "Wikipedia knows the value of what the library catalog forgets." *Cataloging & Classification Quarterly*, *57*(2–3), 166–83. https://doi.org/10.1080/01639374.2019.1597005.

Library of Congress, Policy and Standards Division (2020). "Demographic Group Terms Manual." https://web.archive.org/web/20201029185954/www.loc.gov/aba/publications/FreeLCDGT/freelcdgt.html.

Manske, M. (2015, May 6). Überlistet. *The Whelming*. https://web.archive.org/web/20201024200154/http://magnusmanske.de/wordpress/?p=301.

OCLC Connexion. (2020). *Trombone Shorty by Troy Andrews*. OCLC number 880349715. *OCLC*. https://web.archive.org/web/20201202154550/www.worldcat.org/title/trombone-shorty/oclc/880349715.

OCLC Research (2020). *Libraries Leverage Wikimedia*. *OCLC*. https://web.
archive.org/web/20200927165544/www.oclc.org/research/areas/commu
nity-catalysts/libraries-wikimedia.html

PCC Identity Management Home. (2020). Wikidata Pilot. *PCC*. https://
web.archive.org/web/20201030220924/https://wiki.lyrasis.org/display/
pccidmgt/Wikidata+Pilot

Smith-Yoshimura, K., Altman, M., Conlon, M., Cristán, A. L., Dawson, L., Dun-
ham, J., Hickey, T., Hook, D., Horstmann, W., MacEwan, A., Schreur, P., Smart,
L., Wacker, M., & Woutersen, S. (2014) Registering researchers in author-
ity files. *OCLC Research*. https://web.archive.org/web/20200709205119/
www.oclc.org/content/dam/research/publications/library/2014/
oclcresearch-registering-researchers-2014.pdf.

VuFind 4.1 Milner Library. (2020). Author search for Langston Hughes.
https://web.archive.org/web/20200929150211/https://i-share.carli.illinois.
edu/vf-isu/Author/Home?author=Hughes,+Langston,+1902-1967.

Wikipedia contributors. (2020a). List of African-American writers. *Wikipe-
dia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=
List_of_African-American_writers&oldid=855683364.

Wikipedia contributors. (2020b). Wikipedia:Authority control. *Wikipedia, The
Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Wikipedia:
Authority_control&oldid=990543395

Wikipedia contributors. (2020c). Who is African-American. *Wikipedia, The
Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=African_
Americans&oldid=496452616.

Wikipedia contributors. (2020d). Stanley Bennett Clay. *Wikipedia, The
Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Stanley_
Bennett_Clay&oldid=977742050.

Wikipedia contributors. (2020e). Category:African-American writers. *Wiki-
pedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=
Category:African-American_writers&oldid=969141150.

Willey, E., & Yon, A. (2019). Applying Library of Congress Demographic
Group Characteristics for Creators. *Cataloging & Classification Quarterly*,
57(6), 349–68. https://doi.org/10.1080/01639374.2019.1654054.