VIAFbot and the Integration of Library Data on Wikipedia

This article presents a case study of a project, led by Wikipedians in Residence at OCLC and the British Library, to integrate authority data from the Virtual International Authority File (VIAF) with biographical Wikipedia articles. This linking of data represents an opportunity for libraries to present their traditionally siloed data, such as catalog and authority records, in more openly accessible web platforms. The project successfully added authority data to hundreds of thousands of articles on the English Wikipedia, and is poised to do so on the hundreds of other Wikipedias in other languages. Furthermore, the advent of Wikidata has created opportunities for further analysis and comparison of data from libraries and Wikipedia alike. This project, for example, has already led to insights into gender imbalance both on Wikipedia and in library authority work. We explore the possibility of similar efforts to link other library data, such as classification schemes, in Wikipedia.

by Maximilian Klein and Alex Kyrios

Introduction

Libraries wanting to increase the value and discoverability of their metadata should explore options for making that data available and useful outside of the data silo of the library world. With library users increasingly turning to general search tools like Google before library sources, libraries risk losing relevance by taking a passive role and expecting users to come to them. As profit-driven entities, Google and Amazon don't exhibit such passivity.

Fortunately, Wikipedia is proving a powerful partner for libraries. Some librarians have traditionally regarded Wikipedia with the same sort of skepticism they do Google, as a generalist source that users turn to for "satisficing" in the place of high-quality library information sources. Indeed, Wikipedia has some limitations that make a certain amount of skepticism healthy. But libraries looking to keep their relevance in a digital age need partners, and Wikipedia is well suited to be such a partner (Kyrios 2013).

Unlike Google or Amazon, Wikipedia is run by a nonprofit organization, the Wikimedia Foundation. As of September 2012, the Wikimedia Foundation employed 142 people, compared to over 50,000 for Google and over 91,000 for Amazon. In many ways, the Wikimedia Foundation thinks like a library. It wants to make lots of information available to lots of people, at no cost and for no profit. These are values many librarians share.

There have been many efforts in which libraries have partnered with Wikipedia, generally benefiting both parties. We will focus on a recent effort to bring authority control, a function librarians have long excelled at, to Wikipedia. Library authority work, consisting of creating authorized forms of personal names (as well as names of other entities, such as corporate bodies) helps distinguish individuals with similar names. Wikipedia has its own practices, such as disambiguation pages and redirects, to deal with articles on individuals with the same name. Spearheaded by Wikipedians in Residence at OCLC and the British Library, this project matched Virtual International Authority File (VIAF) identifiers to hundreds of thousands of biographical Wikipedia articles, using a matching algorithm and VIAFbot, an automated Wikipedia account (a "bot").

VIAF

First proposed in 1998, VIAF was created in 2003 as a joint project of OCLC, the Library of

Congress, and the German National Library (Deutsche Nationalbibliothek). In 2007, the National Library of France (Bibliothèque nationale de France) joined, and in 2012, VIAF became an openly accessible OCLC service. VIAF works with institutions to create master authority files. It assigns unique identifiers to each of its records, and also links these records to files maintained by its partner institutions (32 agencies as of June 2013, mostly national libraries) (Murphy 2012). Beginning in summer 2012, VIAF files have also begun to be linked to Wikipedia articles, and vice versa.

Wikipedia and Wikidata

Wikipedia was launched in January 2001, though it hasn't always enjoyed the prominence it does today. Arguably, 2003 was the year the project hit the mainstream, marking its hundred thousandth article, the creation of the Wikimedia Foundation, the first real-life meetup of Wikipedia users, and the creation of the globe logo that the site still uses ten years later. As of June 2013, there are Wikipedias in 285 different languages, though the English Wikipedia is the oldest and largest. This article discusses the English Wikipedia unless otherwise specified.

Readers of Wikipedia will be familiar with the site's encyclopedic articles, but content in these articles is also affected by thousands of behind-the-scenes pages. This project in particular took advantage of such pages known as templates. Templates serve a variety of purposes, but this project mostly used a few templates of structured data which apply to articles about persons. The *Infobox person* template and its derivatives create the boxes of metadata visible towards the top of many personal articles. The *Persondata* template serves a similar function, though it is invisible and has only basic parameters, similar to the fundamentals covered by a MARC authority record. The *Defaultsort* template specifies how an article will be alphabetized. Personal articles typically have a *Defaultsort* specifying a Surname, Given name order. Finally, the *Authority control* template, visible at the bottom of certain personal articles, already existed to link from Wikipedia articles to authority files, though over the course of this project, this template was refined and came into use on hundreds of thousands of more articles.

The project also benefited from the creation of Wikidata, a Wikimedia project that serves as centralized data storage for all Wikipedias. Savvy Wikipedia users may know about Wikimedia Commons, a project which collects media for use across various Wikipedias. Wikidata plays a similar role for data.

Proposal (RfC)

The idea of the VIAFbot project was first proposed on Wikipedia in June 2012 by Wikipedians in Residence at OCLC and the British Library. A description of the project basics received mostly positive reception at the Village Pump, a discussion forum on Wikipedia where new ideas are often proposed. Subsequently, the full proposal was put to a Request for Comment (RfC), a more formalized procedure allowing Wikipedia users to express their support for, or opposition to, the project. Community input soundly endorsed the proposal, with thirty-eight editors in support and only two against (the editors voting against did not oppose authority control itself, but rather the addition of VIAF identifiers at the bottom of articles, which they considered to be cluttered spaces). About a month after the initial proposal, the RfC was closed, with "clear consensus" to proceed with the VIAF integration.

The open, collaborative nature of Wikipedia made it an ideal partner for this project. While the Wikimedia Foundation has some structured leadership, such as its Board of Trustees, these individuals typically have no more editorial authority at Wikipedia than any other user. Thus, a library attempting a similar project need not petition Foundation leadership; an informal consensus among Wikipedia users generally suffices.

The Process

This project was made possible by algorithmic matching between VIAF and English Wikipedia conducted by the VIAF engineering team. Once the VIAF matching algorithm had determined links between the two datasets, the work for VIAFbot became an elaborate copy and paste job. The VIAF matching algorithm is designed to cluster authority file entities. These have properties like *name*, *birth date*, *death date*, and *publications*. Owing to semi-structured data that occurs on English Wikipedia through the use of templates, authority file-compatible schemas could be created from Wikipedia dumps. In particular the templates *Persondata* and *Defaultsort* were most useful. It should be mentioned that matching occurred with one Wikipedia to simplify the matching problem, and English was used in this case for its corpus size, as the largest Wikipedia. However, such matching could be performed on any Wikipedia, or even multiple languages. Data from authority files matches by name and date, and publications were computed into VIAF clusters.

The efficiency of VIAF clusters is improved with the number of contributing files because of the nature of this matching. There can be situations where A matches B and B matches C, but C doesn't match A, yet A,B,C are a cluster. An example of such a case is the VIAF cluster for Bengali author Taslima Narsin. Owing to differing romanizations of her name, many spelling variants exist. The Wikipedia data is close enough to match the entity from Bilbliotheque nationale de France, but not the Union Catalogue of Polish Research Libraries. Yet because the Polish and French entities are close enough to match, VIAF accepts the transitive inference to connect Wikipedia and the Polish entity. More data means better clusters in the VIAF world.

The first VIAFbot operated on English Wikipedia and utilized the pywikipediabot Python-Wikipedia framework. Its starting points were the clusters of the VIAF database that contained an associated English Wikipedia article. There were 269,494 VIAF clusters which had a Wikipedia link. Each of these links was loaded, and the page was scanned to "sanity check" that the page was about an individual. Sanity was accepted if the page had a template that was typically used on articles about people such as *Persondata*, which was the case for 254,678 articles. Of this group 9,034 had preexisting *Authority control* templates. For each "sane" English article the bot followed the German interlanguage link, where possible, to attempt to find any German metadata from previous authority work. (Interlanguage links are links to articles on the same subject in other Wikipedias. They appear on the left hand side of many articles. The Spain article, for example, includes a link to España on the Spanish Wikipedia.)

This interlanguage checking led to 109,087 German articles, of which 92,253 had the *Normdaten* template, the German equivalent of *Authority control*. Of this subset, 74,864 had a VIAF ID (Figure 1). Comparing three possible data sources—English Wikipedia, German Wikipedia, and VIAF—the bot noted how often each disagreed with another. The error rates varied between 10.5% and 15.9% (Figure 3). In the case of any of these conflicts, the article was added to a conflict log and nothing was written to the Wikipedia article. Where there were matches, VIAFbot added a link to the relevant VIAF file in that article's *Authority control* template (and added the template if the article did not already have it).

Figure 1: VIAFbot decision tree

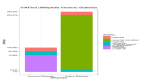

Figure 2: VIAFbot statistics by Wikipedia language as of November 2012

Figure 3: Disagreements among English Wikipedia (Authority control), German Wikipedia (Normdaten), and VIAF.org as of November 2012

Since November 2012, an ongoing period of human correction has followed the bot's initial efforts. As of June 2013, 217 community-led handmade replacements have been made.

After this process was complete, there was interest from other language Wikipedias who wanted VIAFbot to run on their Wikipedias. Users at the Italian Wikipedia took matters into their own hands and migrated what VIAF IDs they could find with interwiki links. In that process they copied about 40,000 identifiers. However, the problem of how to spread this data to all 285 Wikipedias still remained. In times past, the only option would be to write bots to shuffle and translate the templates for each individual Wikipedia. Even then, linked articles would not be guaranteed to have synchronized data—an error could be fixed in a German *Normdaten* but remain in the corresponding English *Authority control*. Fortunately, Wikidata was created to solve such problems, and started providing support in February 2013, while the VIAFbot project was ongoing. As a central datastore of semantic data for groups of pages all related by interlanguage links, its creation in relation to this project was well timed.

Wikidata did not emerge as a live product until after the Italian copy had occurred. With the advent of Wikidata, it was then possible to leverage this multilingual semantic data connector to expand VIAF matching to new Wikipedias. The next iteration of VIAFbot utilized a version of pywikipediabot with only alpha Wikidata support. While the Wikidata version of VIAFbot was custom written, the design pattern has since been incorporated into the script harvest_template.py.

At this juncture, merging the data existing in all of English's *Authority control* templates along with its equivalents in German, French, and Italian was attempted. These languages were selected because they had the highest incidences of authority control templates on their respective Wikipedias.

In order to get the freshest possible data for this operation, the bot actively retrieved live Wikipedia pages rather than working off of a dump. While this was useful because it ensured up-to-date data, the total runtime of the bot extended past three weeks. The starting points for this bot were the lists made by using the special *what links here* bidirectional link function of Wikipedia. When *what links here* is called on a template page, it returns all the pages that transclude (dynamically include) that template, so checking what links to common templates such as *Infobox person* garner very useful, if unwieldy, data sources.

Working off this list of transclusions, each page was transformed using *mwparserfromhell*, a module which converts "wikitext," Wikipedia's markup language, into convenient Python objects representing data fields. Each identifier in the *Authority control* template is jotted down along with a source statement about the language Wikipedia in which it was found.

Then the Wikidata item number was queried from the Wikipedia article. Taking into account any preexisting authority control on Wikidata, the new authority control is either added, or if it matches an existing claim, only the source is added. Wikidata can display seemingly conflicting data as it tries to model provable statements—not truth, reflecting Wikipedia's standard of verifiability, not truth. In total, this process added close to a million statements of authority control from seven sources (Table 1).

Table 1: Authority file statements added to Wikidata (Gray and Klein 2013)

Total number of items with any AC	**417,915**
Authority File	**Amount in Wikidata**
Virtual International Authority File (VIAF)	388,763
Gemeinsame Normdatei (GND)	218,084
International Standard Name Identifier (ISNI)	185,711

Library of Congress Control Number (LCCN)	157,082
Bibliothèque nationale de France (BNF)	15,665
Système Universitaire de Documentation (SUDOC)	12,950
Istituto Centrale per il Catalogo Unico (ICCU)	1,540

Addressing Errors

A sample of 100 random VIAF IDs from Wikidata revealed 98 correct and only two incorrect identifiers. Fortunately, Wikidata was founded with the same open, collaborative model as Wikipedia, so accuracy should improve naturally over time. However, while Wikidata can keep data synchronized across Wikipedias, it cannot do so outside of Wikimedia projects. So when a when a user changes a link from Wikidata into VIAF, the link from VIAF into Wikipedia will not change, leading to a disagreement. To address this issue, it is planned that VIAF will read the inbound links from Wikidata and heal the link discrepancy.

Healing will happen more often in Wikidata than in English Wikipedia because it can be healed by a user speaking any language. There are approximately 85,000 active Wikipedia editors of any language, compared to 30,000 on the English Wikipedia alone.

Web Traffic Impact

Since VIAFbot launched on Wikipedia, VIAF.org has seen a threefold increase in traffic.

Figure 4: VIAF.org traffic statistics

Notable specifically is the increase in traffic coming from English Wikipedia, where the bot first operated. The Wikidata version of VIAFbot started in March and has yet to show much impact, as it is not fully finished. However, the increase in referral traffic is greater than the sum of increase in Wikipedia traffic, so there is also some knock-on effect from secondhand referrals, not from any one particular other referrer. For some perspective, not all traffic to VIAF is referral traffic; from September 2012 until June 2013, VIAF received 577,000 visitors from Google searching, 303,000 from direct visits, and 207,000 from Wikimedia projects.

Analysis Case Study: Sex Data

After the Wikidata import, VIAFbot performed a proof-of-concept auxiliary import that used the connections it had laid.

The property for the sex of a human is one of the most used properties on Wikidata. It is also information that is stored in VIAF, and VIAF's linked data partners' databases. Wikidata was queried using Pywikipedia bot for all the pages that have VIAF as a property. The bot used code such as the following:

```
viaf_property_page = pywikibot.ItemPage(wikidata,
'Property:P214')
#214 is VIAF; stored numerically to avoid language ties.
pages_with_viaf = viaf_property_page.getReferences()
#which returns a generator, that we can cycle through
for page in pages_with_viaf:
#but these instances of class page are just promises,
#until we use a method to get the data
page_parts = page.get()
#and when we do get the page it's returned as dict
claims_list = page_parts['claims']
```

Each item that contains a VIAF ID as a claim can possibly also have a sex as a claim. The VIAF ID string was used to build the URI of the record at VIAF.org. The VIAF.org record provides VIAF's "opinion" on the sex of the entity. That opinion is a result of a behind-the-scenes merge of all the national library files' data. Unfortunately, in this merge not all the data and its provenance is preserved. But because we live in a linked data era, it's possible to follow links out of VIAF into the online databases of the contributing libraries. The Library of Congress system was a good source for this effort because their data model for dealing with complex sex cases is nuanced. Library of Congress will record multiple sexes with applicable dates if they exist, which is a step in the right direction compared to the problematic binary (or trinary, counting those classified as intersex) Wikidata model. Now data could be compared between the Library of Congress and VIAF. Specific information from Library of Congress trumped the merged information on VIAF. Then the single ruling library opinion from that comparison was held to Wikidata. If the opinions matched, the Wikidata claim gained a source where no Wikidata opinion previously existed. In the cases where Wikidata and VIAF disagreed with one another, a list was made for human correction.

Figure 5: Comparison of sex data between VIAF and Wikidata

Of the entire eligible data set of 388,000 Wikidata items with the VIAF property, a subset of 131,650 has both Wikidata sex data and VIAF sex data. Reassuringly, only .2% of this subset disagree. For each of those 311 items in that .2%, human research can right the discrepancy. For instance the Deutsche Nationalbibliothek thinks Nadine Warmuth is male, but Wikidata thinks female (Wikidata is correct). On the other hand, Wikidata thinks that Nguyen Thi Binh is female, but VIAF is correct to suggest otherwise (Klein 2013b). Both Wikidata and library sources make mistakes.

There are also instances without even two sources to conflict with each other. There were 125,781 times when Wikidata had sex data that VIAF did not. This is a case where libraries could glean information from Wikidata. Conversely, Wikidata was pleased to be informed of the 44,526 scenarios where VIAF or LoC had sex information but Wikidata did not.

Lastly, and for perspective, each time sex information was handled, its content was tracked. Of the 257,431 Wikidata items with sex data, 14.7% were female, 85.3% were male, and 0.002% were intersex. The sex data that came from VIAF showed a very similar story at 14.6%, 85.4%, and .006%, respectively, even at a lower sample size of 176,187 (see Figure 6). The closeness between these two measures speaks to shared ingrained biases. Wikipedia, which has been criticized for male bias (Lam et al. 2011), hosts a thoughtful essay on its own systemic bias. This review of sex data suggests that libraries, frequently perceived as progressive institutions, may have some of the same issues. Alternatively, perhaps the sort of systemic bias found on Wikipedia simply reflects the bias of publishing in general.

Figure 6: Composition of Wikidata and VIAF by Sex

This very brief case study of sex data is just one example of the types of research that will be enabled by the connection of the two datasets.

Future Implications

We believe the success of VIAFbot can be replicated with similar sets of structured library data. Integrating authority data with Wikipedia has been called "a milestone in the way library authority data are repurposed for non-traditional uses" (Lovins 2006).This initial project with personal names is just the beginning. VIAF contains other authority files, such as those for corporate names, that could be matched to Wikipedia articles with a similar process. And of course, a huge portion of Wikipedia articles describe neither a person nor a corporate entity. Many of these topical articles could be matched with subject headings from controlled vocabularies such as FAST, LCSH, or MeSH. MeSH terms, in fact, can already be manually linked with a template. As of June 2013, over 6000 medical articles already link to their respective MeSH entries. Articles could also be linked to classification numbers, so Wikipedia readers could follow a link to WorldCat to find library resources for further reading (or especially savvy users could take that number straight to the stacks and browse!). One possible technique is to use the subject classifications of citations of the article to determine the subject classification of the article itself (Klein 2013a). The release of the Dewey Decimal Classification (DDC) system as linked data might allow such a project (Mitchell and Panzer 2013), as might Wikipedia's own articles listing Dewey and LC classes, which link to subject articles (Ayers et al. 2008). Concordances between Wikipedia's classification and Universal Decimal Classification (UDC) have also been explored (Salah et al. 2012).

Additionally, mapping any one of DDC, LCSH, or LCC to Wikipedia articles could aid in mapping the others. Through the use of existing crosswalks, such as those included in the Library of Congress's Classification Web, correlations among these systems could potentially make it easier to link to additional ones. This approach would require refinement, however, or human review, as DDC, LCSH, and LCC do not always match in neat one-to-one relationships. This would advance an interest among Library of Congress personnel in integrating LC vocabularies into linked data environments (Tillett et al. 2011). At least one author has explored the use of Wikipedia data to create a new classification scheme (Yelton 2011). DBpedia, a project to extract structured data from Wikipedia that was founded in 2007, is another potential partner in such efforts (Morsey et al. 2012).

Conclusion

The VIAFbot initiative has connected library authority data with hundreds of thousands of pages on one of the world's most popular websites, increasing the visibility and availability of that data and, by extension, libraries as an institution. The positive reception to the project at Wikipedia affirms the strength of libraries in performing authority control and proves the utility of this work in the era of linked data. The project offers a blueprint for similar efforts to integrate library data with Wikipedia and, perhaps even more importantly, has built good will for the library community with Wikipedia. At a time when many libraries worry about keeping up with information in a digital world, collaborations like this one offer an exciting glimpse of what libraries can still do to help connection their users with high-quality information resources.

References

Ayers P, Matthews C, Yates B. How Wikipedia works: and how you can be a part of it [Internet]. San Francisco (CA): No Starch Press; 2008. Available from: http://archive.org/details/HowWikipediaWorks

Gray A, Klein M. 2013. Wikipedia in the library. Refer 29(2):6-10.

Klein M. 2013a. Wikipedia analytics engine [Internet]. Hanging Together. Available from: http://hangingtogether.org/?p=2452

*Klein M. 2013b. Sex ratios in Wikidata, Wikipedias, and VIAF part 2 [Internet]. Hanging Together. Available from: http://hangingtogether.org/?p=2986

Kyrios A. 2013. Time for libraries to take a fresh look at Wikipedia [Internet]. The Idaho Librarian 63(1). Available from: http://theidaholibrarian.wordpress.com/2013/05/17/time-for-libraries-to-take-a-fresh-look-at-wikipedia/

Lam SK, Uduwage A, Dong Z, Sen S, Musicant DR, Terveen L, Riedl J. 2011. WP:clubhouse?: an exploration of Wikipedia's gender imbalance. In WikiSym '11: Proceedings of the 7th International Symposium on Wikis and Open Collaboration; 2011 Oct 3-5; Mountain View (CA). New York (NY): Association for Computing Machinery. p. 1-10. doi: 10.1145/2038558.2038560

Lovins D. 2006. Cataloging news. Cataloging & Classification Quarterly 42(1):139-147. doi: 10.1300/J104v42n01_10

Mitchell JS, Panzer M. 2013. Dewey linked data: making connections with old friends and new acquaintances. Italian Journal of Library and Information Science 4(1):177-199. doi: 10.4403/jlis.it-5467

Morsey M, Lehmann J, Auer S, Stadler C, Hellmann S. 2012. DBpedia and the live extraction of structured data from Wikipedia. Program: Electronic Library and Information Systems 46(2):157-181. doi: 10.1108/00330331211221828

Murphy B. 2012. Virtual International Authority File service transitions to OCLC; contributing institutions continue to shape direction through VIAF Council [Internet]. 2012 April 4. Dublin (OH): OCLC. Available from: http://www.oclc.org/news/releases/2012/201224.en.html

Salah AA, Gao C, Suchecki K, Scharnhorst A. 2012. Need to categorize: a comparative look at the categories of Universal Decimal Classification system and Wikipedia. Leonardo 45(1):84-85. doi: 10.1162/LEON_a_00344

Tillett B, Dechman L, McLean L. Linking to LCSH and LCC: controlled subject headings and classification systems through the web [Internet]. 2011. Available from: http://conference.ifla.org/past/ifla77/149-tillett-en.pdf

Yelton A. 2011. A simple scheme for book classification using Wikipedia [Internet]. Information Technology and Libraries 30(1): 7-15. Available from: http://ejournals.bc.edu/ojs/index.php/ital/article/view/3040

About the Authors

Maximilian Klein (kleinm@oclc.org) is the Wikipedian in Residence at OCLC.

Alex Kyrios (akyrios@uidaho.edu) is the Metadata and Catalog Librarian at the University of Idaho and a Wikipedia enthusiast.