"Wikidata and Archivists," Archives Association of Ontario Institutional Issues Forum, Archives of Ontario, Toronto, 24 October 2019, http://hdl.handle.net/10315/36898

Wikidata and Archivists

Katrina Cohen-Palacios[1]

About Wikidata and Archives

Hello, everyone. Today, I am going to try and provide a brief, non-technical introduction to Wikidata. Disclaimer: I am by no means an expert on Wikidata. My first edit was approximately a year ago and have only started heavily editing Wikidata only a few months ago.

Wikipedia, a platform familiar to most, is part of the Wikimedia Foundation. The foundation also includes Wikimedia Commons – a platform to upload and share freely usable media files – and Wikidata. Launched in 2012, Wikidata is the youngest Wikimedia project and will celebrate its ninth birthday in a few days. The platform can be described as a "knowledge base" and is essentially a large linked open data set.

In this slide, these screenshots illustrate the differences between a Wikipedia article and a Wikidata item. On the Wikipedia page, the entire article has to be scanned to discover awards received, whereas information can be quickly extracted from the Wikidata item.

The Wikidata item contains bite-sized information that is linked open data because it can be linked with other datasets by following certain standards to be machine-readable (structured vocabulary and unique identifiers) and because is free to use, remix, and publish without copyright restrictions (open).

In our example, information in the sentences about awards received (highlighted on the Wikipedia page) are broken down into semantic triples to convey machine-readable information. The semantic triple formula (subject, predicate, object) follows that of a simple sentence:

- The Toronto Telegram (Q17152683) created (P6241) the Toronto Telegram fonds (Q71342728).
- The Toronto Telegram (Q17152683) employed (P108) John D. Harbron (Q28792876).
- The Toronto Telegram fonds (Q71342728) is part of (P195) the Clara Thomas Archives and Special Collection (Q61467363).
- John D. Harbron (Q28792876) preserved his archives at (P485) the Clara Thomas Archives and Special Collection (Q61467363).

In these examples, the unique identifiers pertain to Wikidata items (Q) and properties (P). Through these simple sentences, we can begin to explore the relationship between two fonds at the Clara Thomas Archives and Special Collections.

Archivists can also add links to their finding aids (or simply indicate that material is at an institution) by using the "archives at" property. This example is from the Edith Fowke item page. On it, there are multiple values added to the archives at property which indicates that the fonds is split between multiple institutions across Canada.

[1] Katrina Cohen-Palacios is a Media Archivist at York University who, over a decade ago, first edited Wikipedia for an undergraduate public history assignment. In 2014, as a graduate student archivist at the University of Toronto, she was asked to insert links to finding aids in Wikipedia articles. Since then, her interest in Wikipedia has continued to grow. With a growing interest in structured data, she started to explore Wikidata and the relationships of holdings within (and beyond) the Clara Thomas Archives and Special Collections.

Add a link to your institution on Wikidata via this property is simple. Click the “add a new statement” option, type “archives at,” and type your institution’s name. If your institution does not pop up, you will have to create an item. You can add more granular levels of detail with qualifiers such as inventory number (P217), described at URL (P973), level of description (P6224), reference URL (P854), and reference retrieval date (P813).

For a list of properties related to archival material, I suggest reviewing the data structure page of the Wikidata archival description project. It explains which properties are best suited for different types of archival items (institutions, collections/fonds, buildings, individuals, etc). Project participants can also propose new properties if one does not exist to fit our needs. The project aims to “create the world’s most comprehensive high-quality database of archival fonds and heritage collections, to represent archival structures within Wikidata where this is deemed useful and to ensure the interlinking between archival finding aids and Wikidata.” The documentation is primarily written in French as the majority of the archivists working on this project are located in Belgium, France, and Switzerland.

Francophone archives have been early adopters of linking their holdings in Wikidata using the archives at property (as seen in this bubble chart illustrating the total number of “archives at” links per country). The good news is that Canada is slowly climbing up the ranks. McGill is currently in the lead and the Clara Thomas Archives & Special Collections just recently imported nearly 200 links to our holdings last Friday. Archivists across the globe continue to adopt Wikidata and in comparing this chart to that of a few months ago. Since then, Belgium surpassed France, the UK jumped ahead of the Netherlands, and Germany and Canada skipped past Bulgaria.

Why contribute…?

Contributing to Wikidata is a way to automate edits, use shortcuts, and ensure information is up to date in Wikipedia. Certain templates are coded to pull information from Wikidata to Wikipedia. For example, in the authority controls are automatically added to the Wikipedia article thanks to a template that pulls the information from the Wikidata item. Using this template on other pages by simply placing the words “authority control” within the double curly brackets, there’s the possibility of adding over 30 different name authorities to the article, if the information exists within Wikidata. Name authorities include the US Congress, NARA, and SNAC (Social Networks and Archival Context) ids. At the moment, SNAC identifiers are references on over 125,083 articles on Wikipedia!

The official website template works in the same way: it pulls information to the article from Wikidata item - saving you time energy and effort. If the information does not exist in Wikidata, the code will remains invisible until the property is populated in Wikidata.

Now, imagine if there was a template that automatically populated links to finding aids by checking information on Wikidata. Typically, archivists manually insert links to finding aids to the external links section of a Wikipedia article. An archival finding aid template could save an incredible amount of time inserting links to finding aids.[2] Not only because it could, for example not only add the link to the Margaret Laurence article but also easily add the finding aid to the nine other language versions of her page.[3]

Now I have very limited experience with Wikimedia Commons (in fact, I have never uploaded an image), but I have discovered that templates in Wikimedia Commons can also pull information from Wikidata. Say, for example, the Archives of Ontario has millions of images in Wikimedia Commons.

[2] Post-presentation update: University of Toronto’s Wikimedian in Residence Alex Jung created a archival records template that references the “archives at” property.
[3] See العربية, Čeština, Deutsch, Français, Հայերեն, Македонски, Română, Русский, and Svenska.

Now imagine that the AO moved its physical location. The information in Wikimedia Commons is now outdated and would have to be manually updated on each and every page of the millions, or it could automatically be updated with one edit in Wikidata because the information in Wikimedia Commons is linked to Wikidata.

So I experimented Wikidata'fying Wikimedia Commons and replaced free text with Wikidata items for a photograph from the Toronto Telegram (see edit history). I added the institution and creator templates to the description:

|institution = {{Institution |wikidata=Q61467363}}
|author = {{Creator|Wikidata=Q17152683}}
|accession number = {{Creator|Wikidata= Q71342728}}

This information will now automatically update on Wikimedia Commons if there are any changes in Wikidata.

As well, by linking the Wikimedia Commons description to Wikidata, additional information about the Clara Thomas Archives and Special Collections was added to the page, including:

- Parent institution: York University Libraries
- Location: Ontario, Canada
- Coordinates: 43° 46′ 21.1″ N, 79° 30′ 19.8″ W Link to OpenStreetMap Link to Google Maps
- Established: 1 July 1970
- Web page: www.library.yorku.ca/web/archives
- Authority control: wikidata:Q61467363: Q61467363 VIAF: 141837883

If you are planning bulk uploads to Wikimedia Commons, I would highly recommend considering a preliminary step: editing and uploading information to Wikidata.[4] Some institutions that have added Wikidata references to their Wikimedia Commons (and have relationships with Wikimedia) include BaNQ, BundesArchiv, and NARA.[5]

If uploading material to Wikimedia Commons does not interest you or your institution as an outreach activity. There are still reasons why you might want to contribute information about your holdings to Wikidata. For example, you can learn more about your holdings through SPARQL queries. In this query, data is extracted from Wikidata to pull the location of CTASC fonds held at other institutions and use their data visualization services to interpret the information as a network graph. Through this graph, I learned that parts of the Inez Elliston fonds is held at the Archives of Ontario, the Edith Fowke fonds is at the Canadian Museum of History, and the Isadora Duncan collection is at the New York Public Library and the Departmental archives of Yvelines. This is only possible because either an archivist or a member of the public had already uploaded information about these holdings to Wikidata.

Another reason to upload material to Wikidata might be related to the AtoM3. The AtoM Foundation states in the AtoM3 Proof of Concept Proposal that: "Data aggregation in portal sites is better

[4] Post-presentation update: Archives may also want to consider creating a Wikidata item for a commons category. For example, the Wikidata item for "images from the Clara Thomas Archives and Special Collections" can generate infobox on the commons page using THIS TEMPLATE and the infobox can generate statistics refering to image usage and pageviews in Wikipedia articles)

[5] Even though the GLAM WikiProject includes archives, incorporating archival material into the wiki-universe can often feel like an exercise of "square peg, round hole." Our information just does not fit the data model. This is most likely because templates are often built for/by librarians and museum professionals. For example, in the description for this file, the archival respoitory (NARA) is listed under "collection." It should also be noted tat NARA created a custom descriptive template which cannot be used by other institutions (as have BaNQ and Bundesarchiv). Archivists may want to work towards creating a single universal archival template that reflects hierarchical arrangement and our descriptive practices?

supported through the use of Linked Data than with traditional hierarchical description. [...] Aggregating data from multiple AtoM instances to portal sites such as Archeion and ArchivesCanada, for example, is a somewhat cumbersome process of importing updated data via CSV and EAD import and replacing existing content. [...] Linked Data, on the other hand, is designed for dynamic querying across multiple information sources, and for assembling results into web interfaces. [...] Linked Data is also better at representing complex relationships and contextual information in archival description. [...] Linked Data schemas, on the other hand, make it possible to capture rich relationships between individuals and communities that intersect with records as creators, custodians and subjects. Rather than a single dominant narrative, a Linked Data approach also allows for a multiplicity of narratives to be captured as statements about a resource, including a multi-provenantial approach to description and arrangement."

In other words, linked data may be the way of the future of archival description and Wikidata is a really easy, low-risk way to experiment and understand the concept and its possibilities. As AtoM3 plans to incorporate linked data, so the transition to the platform may easier if you already understand linked data.

Editing Wikidata

So, if this presentation has convinced you to try Wikidata, there are multiple ways one can edit/add information to the site. To manually edit information, simply find the item through the search bar, click the edit link, type the information, select from the drop-down mention (if a selection does not appear, you must create the item), and then publish.

Another method is Quick Statements, a tool created to batch edit Wikidata. I believe this is how McGill ROARR bulk uploaded "archives at" properties to items about their holdings.[6] I had issues using these methods as there are few controls to disambiguate the names. Note that I only attempted this method once and in a few cases, my data was attached to the wrong individual. Because of this, I had to go into Wikidata and manually remove the data from each item. Based on this experience, I only suggest using this approach with 'clear cut' information (such as languages/countries as demonstrated in the tutorial) which requires little, if any, disambiguation.

My preferred importation method uses Open Refine. I only learned and experimented with this method a week ago to import nearly 200 "archives at" properties for holdings at CTASC using an already compiled list of fonds, identifiers, and URLs from an AtoM migration project.[7] I naively assumed it would be a simple and easy project for a Friday morning. The importation ended up taking me eight hours to learn how to use OpenRefine to reconcile and wrangle the data. If there was a step-by-step tutorial developed for archivists, it may have only taken me a couple of hours.[8] So, I hope to write down and share my documentation once I have learned a little more… stay tuned. But this was the clearest and easiest way. Highly recommended.

Some resources…

[6] See their Access2019 presentation documents (a video recording is also available). In addition to McGill's documentation, I referenced this Wikimedian-in-residence's tutorial to understand how to use Quick Statements in conjunction with the Wikidata Tools add-on for Google Sheets.
[7] See this "OpenRefine Beginners Tutorial = how to process and reconcile datasets" by Emma Carroll at the University of Edinburgh.
[8] Post-presentation update: See Bolliger, Stephanie, Brigitte Bruderlin, Michael Gasser, Julia Lyskawa, Petra Maie, and Lothar Schmitt. "How to Link Your Institution's Collections to Wikidata?: a short manual to a semi-automatic way of using the "archives at" property (P485)." ETH Library, Swiss cultural heritage institutions; Central Library of Zurich; Swiss National Library (January 2020).

As I attempt to understand, learn, and edit Wikidata from an archival perspective, my most commonly referenced resource may have been the #wikidata #p485 hashtags on Twitter (p485 is the identifier for "archives at" property). Through the hashtags, I've witnessed (and joined) discussions with archivists across the globe, primarily Switzerland and France, but includes Australia and the United States. It's also a very rewarding, friendly environment that celebrates small victories.

If you decide to give Wikidata a try, I also recommend reviewing the official Wikidata in Brief and Wikidata Query Service in Brief guides. If you have any questions, please feel free to touch base. I am happy to answer questions and brainstorm possible solutions.

Katrina Cohen-Palacios
kcohenp@yorku.ca
@katrina_cp
wiki: kcohenp