Using Wikidata to Provide Visibility to Women in STEM

Mairelys Lemus-Rojas
Indiana University-Purdue University
Indianapolis, U.S.A.
mlemusro@iupui.edu

Yoo Young Lee
University of Ottawa, Canada
yooyoung.lee@uOttawa.ca

Abstract

Wikidata is an open knowledge base that stores structured linked data. It contains over 58 million items (“Wikidata:Statistics,” n.d.), but its data reveal a noticeable and prevalent gender disparity. In an effort to contribute to the growth and enhancement of women entries in Wikidata, the Indiana University-Purdue University Indianapolis (IUPUI) University Library and the University of Ottawa Library collaborated to embark on pilot projects that broaden the representation and enhance the visibility of women in STEM (Science, Technology, Engineering, and Mathematics). In this article, we share the methods used at both institutions for collecting faculty data, batch ingesting data using external tools, as well as mapping archival data to existing Wikidata properties. We also discuss the challenges faced during the pilot projects.

Keywords: Wikidata; open knowledge; open data; women in STEM; faculty profiles; archival fonds; finding aids

1. Introduction

Wikidata is an open knowledge base that stores structured linked data. Launched on October 29, 2012 by Wikimedia Deutschland, the initial purpose of Wikidata was to serve as the data hub for all Wikimedia projects. Throughout the past seven years, the project has continued growing and evolving to meet the needs of its global community of users. A key feature of Wikidata is its multilinguality. It provides a central repository for all languages to coexist, which is of great benefit to smaller language communities since anyone, regardless of their language skills, can enhance and validate the data (Lemus-Rojas & Pintscher, 2018). Wikidata’s data is published under a CC0 license, which means that it can be consumed and distributed without restrictions. That has facilitated its use by library catalogues, Google Knowledge Graphs, and digital assistant tools like Siri and Alexa (Allison-Cassin & Scott, 2018; Lih, 2018).

Throughout the years, Wikidata has accumulated over 58 million items (“Wikidata:Statistics,” n.d.) but a gender disparity is still very much noticeable in the knowledge base—an issue that its sister project Wikipedia, the free encyclopedia, also faces (“Gender bias on Wikipedia,” 2019). In order to address the gender inequality of women, a few initiatives have emerged in both platforms. The most notable examples are the *WikiProject Women in Red*, which aims to increase the presence of women's issues, women biographies and their works in Wikipedia (“WikiProject Women in Red,” 2019), and the *WikiProject Women,* which has as a main goal to provide proper description for every entry about women in Wikidata ("WikiProject Women," 2019). Despite these efforts, female scientists are still underrepresented in Wikipedia across all levels of scientific achievement (Schellekens, Holstege, & Yasseri, 2019). In an effort to contribute to the growth and enhancement of entries related to women in Wikidata, the Indiana University-Purdue University Indianapolis (IUPUI) University Library and

OPEN ACCESS
This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952141879

the University of Ottawa Library have embarked on pilot projects that broaden the representation and enhance the visibility of women in STEM (Science, Technology, Engineering, and Mathematics). In this article, we share our efforts toward achieving this goal which include using new methods for collecting faculty data, batch ingesting data using external tools, as well as mapping archival fields to existing Wikidata properties. We also discuss some of the challenges we faced when working on these pilot projects at our respective institutions.

2. Faculty Profiles for *Women in STEM* at IUPUI

2.1 Background of Faculty Profiles Project

The IUPUI University Library, located in the IUPUI Campus, has a long history of supporting open access and open knowledge initiatives ("IUPUI University Library Commitment to Open Knowledge," 2019). The library is committed to supporting projects that facilitate information sharing in freely-accessible platforms. As such, a pilot project was conducted in the summer of 2017 to explore the potential of Wikidata and Scholia[1] for IUPUI faculty profiles and to consider the feasibility of implementing a *Scholarly Profiles as Service* model. Scholia, an open source web application that makes live SPARQL (SPARQL Protocol and RDF Query Language) calls to Wikidata, functioned as a scholarly profile generator for the purpose of this project, while Wikidata served as the repository for storing faculty data. The IU Lilly Family School of Philanthropy was chosen for the pilot and Wikidata entries were created for its core faculty members, their co-authors, and some of their publications (Lemus-Rojas & Odell, 2018). The outcomes of this project informed current efforts to provide a presence in Wikidata for IUPUI women faculty—in particular those in STEM fields—and their scholarship.

2.2 *Women in STEM* Project Overview

Wikidata can offer great exposure to faculty publications in the fields of Science, Technology, Engineering, and Mathematics, given that its data can be explored and reused in new services and tools. The *Women in STEM* project offers an opportunity to increase the representation of IUPUI women faculty and their scholarship output in Wikidata, making the data more widely accessible for users. By contributing faculty publications to the ever-growing bibliographic dataset in the knowledge base, contributors are able to establish connections between works. The data can then be explored using Scholia, where citation graphs showing the relationships between works are generated.

In the IUPUI campus, the STEM disciplines are represented in the Purdue School of Science and the Purdue School of Engineering and Technology. For the purpose of this project, it was important to start providing a presence in Wikidata for both schools to be able to connect them with the items to be created for the faculty. All full-time women faculty ranking from assistant to full professor were selected from both schools. The initial process of collecting faculty biographical data from the schools' websites and creating the items was done manually. A different approach to collect data was later taken using Google Sheets, which made it possible to utilize formulas to scrape the data from the web. For this purpose, the IMPORTXML function was used, changing the parameters to fit the needs of the specific website. In this way, the name of the faculty, job title, website URL, and education history were extracted. One challenge faced with this method was that the XML tags in faculty pages were not used consistently, causing the formula to yield errors. Nonetheless, this process was more efficient than having to copy and paste the information manually. After collecting and cleaning the data, properties with constant values were added to the spreadsheet (*instance of* (P31), *sex or gender* (P21), *employer*

[1] Scholia: https://tools.wmflabs.org/scholia/

OPEN ACCESS

This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952141879

(P108), *occupation* (P106), *work location* (P937), *affiliation* (P1416), and *languages spoken, written or signed* (P1412)). Later, the CSV (Comma-Separated Values) file originated from the Google Sheet was formatted using the unique identifiers for the properties in the columns' headers and the identifiers for the items as values in the rows. The file could then be run through the CSV to QuickStatements[2] tool to get the data converted into commands that QuickStatements[3] can read and execute to make batch-edits to Wikidata. For this part of the project, 44 faculty members fit the criteria. As a result, 43 new faculty entries were created and an existing one enhanced in Wikidata. A search can now be performed to get, for instance, all the female faculty affiliated with these two schools.

A preliminary search in Wikidata for publications by the 44 women faculty revealed that some had already been added to the knowledge base. However, the existing entries had the name of the faculty under the *author name string* (P2093) property, which is the default property for authors when the article is added using external tools, such as SourceMetadata. As a result, Scholia was used to get a page listing missing information about the author which is possible to achieve by adding */missing* to the URL. For instance, the page provides a list of author name strings that need to be resolved, any missing co-authors and citing authors, as well as authored works missing topics, all with links to the corresponding tool for the particular work needed. We made use of the link to the Author Disambiguator[4] tool where a list of possible faculty publications that need matching is generated along with the names of all authors and other relevant information. This tool provides a fairly easy and efficient way to disambiguate names and prepares the data to be sent to QuickStatements for batch-editing to Wikidata. Using this method, 173 existing Wikidata article items have been linked to their corresponding faculty. Our goal is to add to Wikidata all publications produced so far by the 44 faculty members selected, and then continue inputting new publications as soon as they are made available.

3. *Canadian Archive of Women in STEM* at the University of Ottawa Library

3.1 Background of *Canadian Archive of Women in STEM* Project

The University of Ottawa Library - Archives and Special Collections initiated a project of Canadian Archive of Women in STEM in collaboration with Library and Archives Canada (LAC) and the International Network of Women Engineers and Scientists - Education and Research Institute (INWES-ERI). The goal of this initiative was to identify archival fonds (finding aids) of women in STEM curated by Canadian institutions and to develop a search portal to facilitate the discovery of these fonds in one central location. As of May 8, 2019, there were 296 archival fonds on Canadian Women in STEM in the portal ("Canadian Archive of Women in STEM," n.d.). These archival fonds contain 10 fields to describe them: media type, call number or archival reference number, STEM field, description, biography, administrative history, date(s), extent, province, and hosting institution. For the purpose of this project, the *WikiProject Archival Description* was identified as a source.

The *WikiProject Archival Description* aims to develop a database of archival fonds and heritage collections in Wikidata and provide reciprocal links between archival finding aids—which offer a detailed description of the collection's content—and Wikidata. Some of the project's tasks include implementation of ontologies regarding the description of archival fonds and data ingestion into Wikidata, and the development of applications using Wikidata's data to improve the user and search experience for archival fonds ("Wikidata:WikiProject Archival Description", n.d.). Given that fonds in the Canadian Archive of Women in STEM portal contain rich archival descriptions, adding its data to

[2] CSV to QuickStatements tool: https://tools.wmflabs.org/ash-django/csv2qs/

[3] QuickStatements tool: https://tools.wmflabs.org/quickstatements/#/

[4] Author Disambiguator tool: https://tools.wmflabs.org/author-disambiguator/author_item.php?

OPEN ACCESS
This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952141879

Wikidata has the potential to serve as a good model for the *WikiProject Archival Description.* The *Canadian Archives of Women in STEM* project focused on developing a mapping between fields from the Canadian Archives of Women in STEM portal and existing Wikidata properties.

Although the portal of Canadian Archive of Women in STEM functions as a searchable index to discover fonds on women and organizations in STEM across Canada, it supports neither an API (Application Programming Interface) to interact with other applications or exchange data between systems, nor tools to extract or visualize datasets. In the era of linked data and big data, representation in Wikidata could improve findability and searchability of the fonds about Canadian women in STEM, rather than having an isolated stand-alone portal.

3.2 Defining Data Structure for Archival Fonds in Wikidata

There are various suggestions to define data structure for archival fonds in Wikidata. For example, *WikiProject Archival Description* suggests the creation of items for the archival fonds and the building in which they are located, as well as the institution that maintains them. ("Wikidata:WikiProject Archival Description", n.d.). Another approach has been to use the *archives at* (P485) property to establish the connection between the entries for the corporate body, person or family to the institution hosting the archival records (Lemus-Rojas & Pintscher, 2018). The goal of the *Canadian Archives of Women in STEM* project is to create/enhance items in Wikidata to represent women/organizations and then connect them to both their corresponding items for the archival fonds and the institutions hosting such fonds.

Using WikidataR[5]—an API library accessible from R, the programming language—the number of Wikidata items was identified for Canadian women/organizations in STEM and the hosting institutions that already exist in the knowledge base. Out of 296 archival fonds, there are 67 women/organizations and 279 hosting institutions represented. Some of the items for women and organizations are more complete than others, which provides an opportunity to enhance those in need and to use those that are more fully described as examples. In contrast, existing items for archival fonds are scarce. This was a motivator for mapping archival fonds to Wikidata properties following the data structure guidelines defined by the *WikiProject Archival Description*, which resulted in the creation of a template that is now being used to support the project. The women/organizations are connected to the hosting institution using the *archives at* (P485) property and linked to their corresponding archival fonds using the *statement is subject of* (P805) as a qualifier of the *archives at* property. The archival fonds are connected to the women/organizations using the *collection creator* (P6242) and *main subject* (P921) and hosting institution using the *maintained by* (P126) property.

3.3 Contributing Data about Archival Fonds to Wikidata

Once the Wikidata properties for *Canadian Archive of Women in STEM* were defined, a new item was created to represent the project. When creating entries for the archival fonds, the property *instance of* (P31) with *archives* (Q56648173) as value, and property *part of* (P361) with *Canadian Archive of Women in STEM* (Q63647303) as value, were used. Given the fact that some items for hosting institutions and women/organizations were already present in Wikidata, the emphasis was on creating entries for their corresponding archival fonds and establishing the connections among them. As of May 8, 2019, there were six Wikidata items representing archival fonds which were manually created and linked to the items for the women and hosting institutions.

[5] WikidataR: https://cran.r-project.org/web/packages/WikidataR/vignettes/Introduction.html

OPEN ACCESS
This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952141879

For this project, most of the time and efforts were dedicated to learning and understanding how Wikidata works and how different concepts can be modeled, in particular those related to archival fonds. One of the challenges faced was how to deal with box sizes in Wikidata when using the *collection or exhibition size* (P1436) property. Traditionally, in Canadian archival practices, the physical description of the archival fonds is entered using the measurement for the width of the box (i.e. 7.5 cm of textual records). But this format is not validated in Wikidata when added as a value to this property. Instead, the number of boxes containing the archival fonds is what should be added. Therefore, there is a need for converting from centimeters to the number of boxes for all archival fonds. During the summer of 2019, the library is expecting to add 67 items using QuickStatements to describe archival fonds for which an item for the creator already exist in Wikidata. The remaining archival fonds, women/organizations, and hosting institutions will be added by the end of the year.

4. Conclusions

Although we took different approaches, our projects share the same theme—Women in STEM—and the ultimate goal of contributing data to Wikidata. As part of the process, we found that existing Wikidata properties were sufficient for the type of data with which we were working. We both recognized that Wikidata has great potential for improving discoverability, searchability and findability of resources in open infrastructure environments. Therefore, contributing to the knowledge base is critical to enhance the visibility of women in STEM. In order to achieve this goal, the Wikidata entries for IUPUI women faculty in STEM and their scholarly output, as well as the Canadian women/organizations in STEM and their archival fonds, were created using various tools to help streamline our processes. The methods used in these projects can be easily adapted to fit the needs of other institutions wanting to provide exposure to faculty and archival fonds in Wikidata. By contributing data about women in STEM to an open platform like Wikidata, we are making a small but important contribution toward closing the gender gap. Our hope is for other institutions to join us in this effort.

References

Allison-Cassin, S., & Scott, D. (2018, May 4). Wikidata: a platform for your library's linked open data. Code4ib Journal(40). Retrieved from https://journal.code4lib.org/articles/13424

Canadian Archive of Women in STEM. (n.d.). Retrieved on May 8, 2019 from https://biblio.uottawa.ca/en/women-in-stem/about

Gender bias on Wikipedia (2019). Retrieved on August 9, 2019 from https://en.wikipedia.org/wiki/Gender-bias_on_Wikipedia

IUPUI University Library Commitment to Open Knowledge. (2019, May 13). Retrieved from https://web.archive.org/web/20190515174925/http://www.ulib.iupui.edu/digitalscholarship/openknowledge

Lemus-Rojas, M., & Odell, J. (2018). Creating Structured Linked Data to Generate Scholarly Profiles: A Pilot Project using Wikidata and Scholia. Journal of Librarianship and Scholarly Communication, 6(1), eP2272. https://doi.org/10.7710/2162-3309.2272

Lemus-Rojas, M., & Pintscher, L. (2018). Wikidata and libraries: Facilitating open knowledge. In M. Proffitt (Ed.), *Leveraging Wikipedia: Connecting communities of knowledge* (pp. 143–158). Chicago: ALA Editions; pre-print version available at http://hdl.handle.net/1805/16690

Lih, A. (2018). *Wikidata 201: An introduction for the citation-oriented* [PowerPoint slides]. Retrieved from https://docs.google.com/presentation/d/1erJU5WiEnIGMVF_VRSJ9P6kit1iPuQMmaJ8j6817g5k/edit#slide=id.g1f49679681_0_183

Schellekens, M., Holstege, F., & Yasseri, T. (2019). Female scholars need to achieve more for equal public recognition. https://arxiv.org/abs/1904.06310

Wikidata:Statistics. (n.d.). Retrieved on May 16, 2019 from https://www.wikidata.org/wiki/Wikidata:Main_Page

OPEN ACCESS

This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

Wikidata:WikiProject Archival Description. (n.d.). Retrieved on February 27, 2019 from https://www.wikidata.org/wiki/Wikidata:Main_Page

WikiProject Women. (2019, May 13). Retrieved from https://www.wikidata.org/wiki/Wikidata:WikiProject_Women

WikiProject Women in Red. (2019, May 13). Retrieved from https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red

OPEN ACCESS

This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952141879