SPARQL for LIS Analytics: Exploring Sex and Gender Representation within Contributions made by PCC Wikidata Pilot Participants

Darnelle Melvin
University Libraries
University of Nevada, Las Vegas

Keywords: Wikidata, SPARQL, graph analytics, gender representation, sexual representation

Introduction

Exploring sex and gender representation in public knowledge graphs presents an opportunity to better understand societal dynamics through an inclusive lens that transcends traditional gender norms. However, with the emergence of semantic technologies and a vast wealth of structured open data available on Wikidata (Vrandečić & Krötzsch, 2014), the gallery, library, archives, museums and special collection (GLAMS) communities are presented with an opportunity to explore sex and gender in new and insightful ways. SPARQL (Feigenbaum et al., 2013), with its ability to query and manipulate linked data, emerges as a compelling instrument for such exploration. In this paper, we present a method which utilizes SPARQL queries and the Wikidata Query Service[1] to investigate the complexities of sex and gender within contributions made by the participants of the Program for Cooperative Cataloging (PCC)[2] Wikidata Pilot (WikiProject PCC Wikidata Pilot - Wikidata, 2020). Through this innovative approach, we seek to shed light on the complexities of gender in the cultural heritage community and pave the way for future research and exploration in this critical area.

Background

In 2020, the PCC Task Group on Identity Management for the Name Authority Cooperative Program (NACO) announced a Wikidata Pilot which ran from September 2020 through December 2021 (PCC Task Group on Identity Management, 2020). During this pilot, 75 PCC member organizations and non-members experimented with identity management by contributing library collection data to Wikidata (WikiProject PCC Wikidata Pilot/Participants, 2020). Each participating organization contributed at its own pace, with projects focusing on diverse themes, often highlighting regional history from archival collections, oral histories, and notable figures (Bergland et al., 2023; Melvin & Lampert, 2023; Zhang et al., 2023).

Methods

The investigation begins with the generation of a list of Wikidata Q-numbers representing the project pages of the pilot participants to assess their contributions. The initial SPARQL query focusses around the Wikidata item 'WikiProject PCC Wikidata Pilot' (wd:Q102157715), which selects objects based on the property 'has part(s)' (wdt:P527).[3] The resulting list comprises 38 pilot pages; however, it should be noted that not all participants have created pages. The query results are saved as a .csv for later use.

Counting Human Contributions: Next, the total human contributions by pilot participants are counted. Figure 1 presents the SPARQL query that selects humans matching the property 'on focus list of Wikimedia project' (wdt:P5008), where each participant's object value is represented. Often, items created or edited during the pilot were tagged with this property to indicate relevance to a Wikimedia project. The Q-numbers obtained from the first query are substituted into a list of values nested within the second query. The resulting data is recorded for analysis.

Counting Human Contributions with Sex or Gender (P21) Values: Finally, we count the number of human contributions with a value assigned for the property 'sex or gender' (wdt:P21) by PCC participants. Figure 2 illustrates the query selecting humans matching the 'on focus list of Wikimedia project' (wdt:P5008), identified as 'human' (wd:Q5), and having a 'date of birth' (wdt:P569) value. The date of birth value is then filtered for dates equal to or greater than January 1, 1950, to ensure that the results reflect the current population and minimize inclusion of historical figures. An additional filter is used to exclude blank nodes. Additionally, the results are grouped by 'sex or gender' (wdt:P21) and ordered by number in descending order. Finally, the results are saved for analysis.

Results

[1] https://query.wikidata.org/
[2] https://www.loc.gov/aba/pcc/
[3] https://w.wiki/9iVZ

At the time of this writing, results reveal that the total number of human contributions by the pilot participants surveyed totals 78,098 items, out of which 35,214 (45.09%) do not have a 'sex or gender' (wdt:P21) value assigned. The distribution of sex or gender assignment among pilot contributions leans heavily towards sexual identities,[4] with 37,794 (48.39%) items assigned 'male' (wd: Q6581097) and 5,086 (6.51%) assigned 'female' (wd: Q6581072). For a complete distribution, see Table 1.

Discussion and Conclusions

The lack of sex or gender assignment in this subset aligns with previous studies on Wikidata gender representation (Klein et al., 2016; Zhang et al., 2021). However, despite efforts to address the gender gap (Address the gender gap/Initiatives - Meta, 2015), two scenarios may explain the high percentage of omission for sex and gender (wdt:P21) property values. Firstly, non-binary individuals may be underrepresented or unidentified due to identity misrepresentation, privacy, or safety concerns (Thompson, 2016; Melvin & Lampert, 2023). Secondly, the 2022 PCC recommendation to omit gender information from name authority records may also have influenced the results (Billey et al., 2022). These issues require further investigation and highlights the need for a more equitable sexual and gender identity representation within Wikidata.

Our use of SPARQL for gender representation analysis in Wikidata highlights its potential as a useful analytics tool. While this study provides valuable insights, it is not exhaustive. Instead, it demonstrates SPARQL's efficacy for conducting analysis on live knowledge graphs for gender representation studies in digital environments.

References

Address the gender gap/Initiatives - Meta. (2015, March 2). wikidata.org. Retrieved April 11, 2024, from https://meta.wikimedia.org/wiki/Address_the_gender_gap/Initiatives

Bergland, K., Dezelar-T, C., & Harrington, P. (2023). The Minnesota Hip-Hop Collection and Wikidata: Practical and ethical challenges for linked data creators. In A. Provo, K. Burlingame, B. M. Watson, (Eds.) *Ethics in Linked Data* (1st ed., pp. 265-293). Library Juice Press.

Billey, A., Colbert, J., Haugen, M. Hostage, J., Ilik, V., Sack, N., & Schiff, A., L. (2022). *Revised report on recording gender in personal name authority records*. PCC Ad Hoc Task Group on Recording Gender in Personal Name Authority Records. https://www.loc.gov/aba/pcc/documents/gender-in-NARs-revised-report.pdf

Feigenbaum, L., Williams, G. T., Cark, K. G., & Torres, E. (2013). SPARQL 1.1 Protocol (W3C Recommendation). https://www.w3.org/TR/sparql11-protocol/

Klein, M., Gupta, H., Rai, V., Konieczny, P., & Zhu, H. (2016). Monitoring the gender gap with Wikidata human gender indicators. *In Proceedings of the 12th International Symposium on Open Collaboration (OpenSym 2016)*, 1–9. https://doi.org/10.1145/2957792.2957798

Melvin, D., & Lampert, C. (2023). Ethical explorations using Wikidata and Wikidata tools to expose underrepresented special collection materials. In A. Provo, K. Burlingame, B. M. Watson, (Eds.) *Ethics in Linked Data* (1st ed., pp. 411-435). Library Juice Press.

PCC Task Group on Identity Management. (2020, June). *Wikidata pilot - PCC Identity Management - LYRASIS wiki*. https://wiki.lyrasis.org/display/pccidmgt/Wikidata+Pilot

Thompson, K. J. (2016). More Than a Name: A Content Analysis of Name Authority Records for Authors Who Self-Identify as Trans. *Library Resources & Technical Services, 60*(3), 140–155. https://doi.org/10.5860/lrts.60n3.140

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78-85. http://dx.doi.org/10.1145/2629489

Wikidata:WikiProject PCC Wikidata Pilot - Wikidata. (2020, August 27). wikidata.org. Retrieved April 8, 2024, from https://www.wikidata.org/wiki/Wikidata:WikiProject_PCC_Wikidata_Pilot

Wikidata:WikiProject PCC Wikidata Pilot/Participants - Wikidata. (2020, September 30). wikidata.org. Retrieved April 8, 2024, from https://www.wikidata.org/wiki/Wikidata:WikiProject_PCC_Wikidata_Pilot/Participants

Zhang, C. C., & Terveen, L. (2021). Quantifying the gap: A case study of Wikidata gender disparities. *In 17th International Symposium on Open Collaboration (OpenSym 2021)*, 1-12. https://doi.org/10.1145/3479986.3479992

Zhang, E., Biswas, P., & Dagher, I. (2023). Reflections on the PCC Wikidata pilot at UCLA library: Undertaking the PCC learning objectives. *Cataloging and Classification Quarterly*, *61*(7-8), 735-772. https://doi.org/10.1080/01639374.2023.2269416

[4] For details on sex and gender identities within the Wikidata ontology see, https://wigedi.com/model

```
#COUNT OF HUMAN (wd:Q5) CONTRIBUTIONS BY PCC PILOT PARTICIPANTS
SELECT (COUNT(?person) AS ?personCount)
WHERE
{
 VALUES (?participant)
 {
  (wd:Q100202113)(wd:Q104813508)(wd:Q100975219)(wd:Q105757729)
  (wd:Q101109490)(wd:Q106675203)(wd:Q100146182)(wd:Q100998622)
  (wd:Q100152473)(wd:Q104694359)(wd:Q105936481)(wd:Q103827632)
  (wd:Q100748830)(wd:Q107300606)(wd:Q104822025)(wd:Q105620684)
  (wd:Q102024112)(wd:Q104665671)(wd:Q101440912)(wd:Q100325511)
  (wd:Q100363299)(wd:Q106518324)(wd:Q107077160)(wd:Q103505599)
  (wd:Q105412680)(wd:Q106908970)(wd:Q100999455)(wd:Q100424907)
  (wd:Q105996609)(wd:Q105395197)(wd:Q103138537)(wd:Q103136167)
  (wd:Q100202113)(wd:Q100235802)(wd:Q100153988)(wd:Q105514917)
  (wd:Q107541094)(wd:Q98970039)(wd:Q106299887)
 }
 ?person wdt:P5008 ?participant .
 ?person wdt:P31 wd:Q5 .
 SERVICE wikibase:label {bd:serviceParam wikibase:language "en" . }
}
```

Figure 1: SPARQL query to count total human contributions by PCC participants.[5]

```
#COUNT ALL HUMAN CONTRIBUTIONS WITH  SEX OR GENDER (wdt:P21) VALUE ASSIGNED BY PCC PARTICIPANT
SELECT ?sexOrGender ?sexOrGenderLabel ?number
WHERE
{
 {
  SELECT ?sexOrGender (COUNT(DISTINCT ?item) AS ?number)
  WHERE
  {
   VALUES (?participant)
   {
    (wd:Q100202113)(wd:Q104813508)(wd:Q100975219)(wd:Q105757729)
    (wd:Q101109490)(wd:Q106675203)(wd:Q100146182)(wd:Q100998622)
    (wd:Q100152473)(wd:Q104694359)(wd:Q105936481)(wd:Q103827632)
    (wd:Q100748830)(wd:Q107300606)(wd:Q104822025)(wd:Q105620684)
    (wd:Q102024112)(wd:Q104665671)(wd:Q101440912)(wd:Q100325511)
    (wd:Q100363299)(wd:Q106518324)(wd:Q107077160)(wd:Q103505599)
    (wd:Q105412680)(wd:Q106908970)(wd:Q100999455)(wd:Q100424907)
    (wd:Q105996609)(wd:Q105395197)(wd:Q103138537)(wd:Q103136167)
    (wd:Q100202113)(wd:Q100235802)(wd:Q100153988)(wd:Q105514917)
    (wd:Q107541094)(wd:Q98970039)(wd:Q106299887)
   }
   ?item wdt:P5008 ?participant .
   ?item wdt:P31 wd:Q5 ; wdt:P21 ?sexOrGender .
   ?item wdt:P569 ?dob .
   FILTER("1950-01-01"^^xsd:dateTime >= ?dob )
   FILTER (!wikibase:isSomeValue(?sexOrGender))
  }GROUP BY ?sexOrGender
 }SERVICE wikibase:label {bd:serviceParam wikibase:language "en" . }
}ORDER BY DESC (?number)
```

Figure 2: SPARQL query to count human contributions by PCC participants with 'sex or gender' values.[6]

Sex or Gender	Contributions	Percent of Gender Contributions
male	37,794	48.39304464
not gendered	35,214	45.0895029
female	5,086	6.512330661
non-binary	1	0.001280443
genderqueer	1	0.001280443
assigned female at birth	1	0.001280443
trans woman	1	0.001280443
Total	**78,098**	**100%**

Table 1: Distribution of sex or gender identities amongst PCC Wikidata pilot participants.[7]

[5] https://w.wiki/AE9B
[6] https://w.wiki/AEA5
[7] The "not gendered" value was derived by subtracting the sum of other categories from the total.