How does Wikidata shape gender identities? Initial findings and developments from the WiGeDi project

Daniele Metilli
University College London

Beatrice Melis
University of Pisa

Chiara Paolini
KU Leuven

Marta Fioravanti
oio.studio

Abstract

The Wikidata Gender Diversity (WiGeDi) project, funded through the Wikimedia Research Fund program, is studying gender diversity in Wikidata, focusing in particular on marginalized gender identities. It will start by examining how the current Wikidata ontology model represents gender, and the extent to which this representation is fair and inclusive. Afterwards, the research will focus on the data stored in the knowledge base to gather insights and identify possible gaps. Finally, it will look at how the community has handled the move towards the inclusion of a wider spectrum of gender identities.

Keywords: Wikidata; gender identities; ontology model; corpus analysis; data ethics.

Introduction

Wikidata is one of the largest general-purpose knowledge bases ever created, and a collaborative project that allows the editing of knowledge — and even the data model itself — by a broad community of users (Vrandečić and Krötzsch, 2014). Data from Wikidata is applied in countless other projects, including Wikipedia.

As in any crowdsourced project, data from Wikidata is subject to biases and gaps, and a critical approach is required when reusing it. However, how these biases are formed and the processes through which the Wikidata community aggregates, compiles and models the data, are currently understudied. It is particularly important to focus on those aspects that may lead to discrimination of marginalized communities, and one of these is gender.

The Wikidata Gender Diversity (WiGeDi) project (Metilli and Paolini, in press) aims to study gender diversity in Wikidata in a wide and comprehensive way, focusing especially on marginalized gender identities such as the trans and non-binary communities. Digital projects have struggled to cope with the wider societal acceptance of the fact that gender is not binary, and in many cases, they have perpetuated — or even amplified — the misgendering and erasure of trans and non-binary people that has occurred in society throughout history. Wikidata is no exception, as it has sometimes become a battleground between inclusive views and narrow views of gender.

In WiGeDi, we are studying gender diversity in Wikidata through the lens of critical data studies (Kitchin and Lauriault, 2014; Iliadis and Russo, 2016), looking at how users have developed a shared understanding and conceptualisation of gender. We explicitly adopt a feminist, intersectional, and queer perspective, viewing gender as a social construct (Butler, 1999) and rejecting the traditional Western view of gender as a binary (DeVun, 2020). Modelling biographical data, and in particular sensitive personal data such as gender, is a "matter of care" (de La Bellacasa, 2017) that requires ethical consideration. Wikidata's shared modelling of gender was hindered by several missteps (e.g., lack of clear ethical guidelines, overuse of automation). Moreover, the users' enthusiasm for open knowledge and the goal of achieving data completeness (i.e., describing the gender of every person) obscured the fundamental questions that loomed in the background: What is gender? Do we need to model it? What is the best way to do it (and best for whom)?

Not many studies have focused on gender in Wikidata so far, and most research has approached the gender gap from a binary perspective. The first scholars to address this topic were (Klein et al., 2016) and (Konieczny and Klein, 2018), while a more recent study was conducted by (Zhang and Terveen, 2021). The topic has been studied more extensively in Wikipedia, e.g. (Beytía and Wagner, 2022; Field et al., 2022; **?**). We take into account the previous studies and the insights they provided, but our perspective is more focused on gender diversity.

Our project studies adopts three complementary perspectives: *model*, *data*, and *community*. We believe that only by answering all three questions it will be possible to obtain a comprehensive overview of gender diversity in Wikidata. Based on our research, we will build a web application to present the results in a user-friendly way.

Model

To begin with, we are investigating how the current Wikidata ontology model represents gender. We attempt to understand the extent to which this representation is fair and inclusive, and how it supports the representation of a wider spectrum of identities by looking at how the

Wikidata ontology has evolved over time. On a technical plane, we are looking at the classes and properties that make up the Wikidata ontology, and how they connect to each other. We have already mapped the current version of the model and performed a preliminary qualitative analysis (Metilli and Paolini, in press); the current aim is to expand this initial work by investigating the limitations of the model and its evolution throughout the history of Wikidata. One result of this work will be the Wikidata Gender Timeline: a tool to narrate how gender has been modeled since the beginning of the project, contextualized with information about real-world events.

Data

Beyond looking at the model, we are also analysing the data stored in the knowledge base in a quantitative way, to gather insights and identify possible gaps with regard to diverse and marginalized gender identities. For example, we are comparing the percentage of trans and non-binary people represented in Wikidata to the prevalence of these populations in society, and our initial findings show a clear under-representation of these populations in Wikidata (Metilli and Paolini, in press). We are also looking at contextual data such as geographic provenance, date of birth, occupation and other relevant data points such as statement sourcing, completeness of the descriptions, multilingual labeling, number of linked Wikipedia biographies, etc. We have started the development of a Wikidata Gender Dashboard that will show real-time statistics and visualize interesting findings.

Community

Finally, we are looking at how the community handles the move towards the inclusion of a wider spectrum of gender identities. Gender representation is often intrinsically connected to language, and this is especially relevant in a multilingual project such as Wikidata. Therefore, we analyze user discussions about the topic of gender identities through computational linguistics methods, using the WiGeTa (Wikidata Gender Talk) corpus, a collection of Wikidata discussions in English[1] on gender-related topics discussed among the users in a ten-year span (Metilli and Paolini, in press). The investigation of the collected discussions is carried out using ATLAS.ti Windows (Version 22.0.6.0), following a critical discourse analysis approach. On the quantitative side, we employ a topic modeling analysis using Latent Dirichlet Allocation (LDA). The two studies help us understand how users would discuss issues related to gender over ten years, letting emerge a temporal evolution of the discussions about this topic.

[1]We are aware that focusing only on English-language discussions is a significant limitation. We are considering widening the scope to other languages as future work.

Conclusions

In the WiGeDi project, we are looking at how the Wikidata community has approached the complex issue of representing gender. The combination of qualitative and quantitative analyses allows for a wide perspective on different aspects, from model to data to user discussions, allowing in-depth research on the community's shared understanding of gender and opening up further studies on the impact of Wikidata's approach to gender on Wikimedia and third-party projects. We hope that our research will help develop more inclusive policies towards gender-diverse people in Wikidata and beyond.

References

[Beytía and Wagner2022] Pablo Beytía and Claudia Wagner. 2022. Visibility layers: A framework for facing the complexity of the gender gap in Wikipedia content. *Internet Policy Review*, 11(1).

[Butler1999] Judith Butler. 1999. *Gender trouble: Feminism and the subversion of identity*. Routledge.

[de La Bellacasa2017] Maria Puig de La Bellacasa. 2017. *Matters of Care*. University of Minnesota Press.

[DeVun2020] Leah DeVun. 2020. *The Shape of Sex*. Columbia University Press.

[Field et al.2022] Anjalie Field, Chan Young Park, Kevin Z Lin, and Yulia Tsvetkov. 2022. Controlled analyses of social biases in Wikipedia bios. In *Proceedings of ACM Web 2022*, pages 2624–2635.

[Iliadis and Russo2016] Andrew Iliadis and Federica Russo. 2016. Critical data studies: An introduction. *Big Data & Society*, 3(2).

[Kitchin and Lauriault2014] Rob Kitchin and Tracey Lauriault. 2014. Towards critical data studies. In *Thinking Big Data in Geography*. University of Nebraska Press.

[Klein et al.2016] Maximilian Klein, Harsh Gupta, Vivek Rai, Piotr Konieczny, and Haiyi Zhu. 2016. Monitoring the gender gap with Wikidata human gender indicators. In *Proceedings of OpenSym*, pages 1–9.

[Konieczny and Klein2018] Piotr Konieczny and Maximilian Klein. 2018. Gender gap through time and space. *New Media & Society*, 20(12):4608–4633.

[Metilli and Paoliniin press] Daniele Metilli and Chiara Paolini. in press. Non-binary gender representation in Wikidata. In *Ethics in Linked Data*. Litwin Books. `https://wigedi.com/chapter.pdf`.

[Vrandečić and Krötzsch2014] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

[Zhang and Terveen2021] Charles Chuankai Zhang and Loren Terveen. 2021. Quantifying the gap: a case study of Wikidata gender disparities. In *Proceedings of OpenSym*, pages 1–12.