# Assessing Sociocultural Disparities in Wikidata between Knowers and Ways-of-Knowing

Loren Koenig
(independent researcher)

## Abstract

Wikidata has been critiqued for concentrating the design and development of important parts of its ontology in the hands of too few. If this is true, Wikidata items and their statements contributed by one sociocultural group may have to rely properties and items contributed by a significantly disjoint other group, calling into question the true representation of the first group's knowledge, seeing that their way-of-knowing is not used in the knowledge representation. This project addresses that question by an analysis of indications of sociocultural affiliation in contributors' user pages.

## Introduction

Several authors and community members have pointed to potential bias Wikidata, and in particular in how Wikidata encodes ways-of-knowing into properties and the upper ontology (e.g., Allison-Cassin 2022; Maitreyi 2023). This may mean that not every Wikidata contributor can express their knowledge in its original way-of-knowing—we formulate the following (intentionally provocative) hypothesis:

**RH:** Wikidata item contributors do not share the sociocultural background of the contributors of properties, upper ontology, etc. that they use.

Or, more pointedly: "(Many) Wikidata 'content' contributors are forced to contribute in a 'language' that is not 'their own'."

This hypothesis is underdefined from a statistics standpoint. The proposed research is largely exploratory, to determine if there are operationalizations of the hypothesis that are answerable with available on-wiki data.

The questions to be addressed by this research are as follows:

**RQ1:** What measures for the sociocultural background of Wikidata users can be constructed based on the readily available, machine-tractable on-wiki indications of sociocultural affinity? (e.g., language as a proxy for culture, 'sociocultural world region', …)
**RQ2:** Do those measures permit to determine potential discrepancies in sociocultural context between item editors or 'upper' editors relevant to the respective items?
**RQ3:** If not, what kind of research would be required?

**Date**:
Start date: June 1, 2024
End date: February 28, 2025

## Related work

The sociology of Wikidata contributors has been researched widely, but mainly to identify 'types' or 'roles' of contributors or typical patterns of contributions (Müller-Birn et al. 2015; Piscopo and Simperl 2018; Sarasua et al. 2019; Zhang et al. 2022). Some research has investigated methods for (and risks of) determining sociographic attributes of Wikipedia contributors (Brückner, Lemmerich, and Strohmaier 2021).

## Methods

This research will use SPARQL queries or similar methods to retrieve properties and items of interest ('toward the upper ontology') to a given Wikidata item. Wikimedia Cloud Services or data dumps will be used to retrieve edit histories and thus all contributors. Various TBD methods (cf. RQ1) similar to those by Brückner, Lemmerich, and Strohmaier (2021) will provide information on the sociographic background of those contributors, which will be

aggregated and compared for the given item vs. its 'upper' items and properties. A random sample of all Wikidata items will be analyzed in this fashion in order to answer RQ2. If needed, RQ3 will be addressed with exploratory literature research.

## Expected output

If RQ2 can be affirmed, this project will provide:

- a method to compare sociocultural context of a given Wikidata items with that of the items and properties used in its statements
- confirmation or rejection of RH

In any case, the project will produce

- a research paper
- future research directions

These outputs should be particularly valuable to communities who wish to assess whether their representation in the item space on Wikidata is matched by equal representation in the properties and upper ontology representing their way-of-knowing.

## Risks

This research carries certain privacy risks to Wikidata contributors. For a discussion, see Brückner, Lemmerich, and Strohmaier (2021). Following their suggestion, inferred sociographic attributes of Wikidata contributors should only be stored and be made available on the aggregate level (i.e., across all contributors to a given item or property).

## Community impact plan

In the conduct of this research, some continuous collaboration with the greater Wikimedia research community will be mutually beneficial. This may occur at hackathons, on-wiki, etc. Apart from this, the at least one-time presentation of methods and results in a suitable venue (conference, journal) should be expected to stimulate future research in this area.

## Evaluation

An affirmative answer to RQ2 and thus confirmation or rejection of RH would be a major success. Otherwise, the identification of future research directions, and the Wikimedia research and developer community being able to make use of even partial results and methods would count as some success.

## Budget

[USD]

| 40,000 total; breakdown is as follows: |
|---|
| 32,500 researcher/developer remuneration |
| 3,500 open-access publishing fee |
| 2,000 conference participation |
| 2,000 potential 3rd-party computing-infrastructure |

## Prior contributions

I have over the years facilitated many discussions and given talks about Wikidata at events organized by Wikimedia NYC. More recently, with Provo et al. (2021), I co-facilitated a critical discussion among librarians about sociotechnological questions regarding Wikidata and its use in biliographic contexts and library instruction. At WikidataCon 2021 I presented a graphical formalism to discuss the question of whether all knowledges are equally representable, and all ways-of-knowing equally represented, in Wikidata (Koenig 2021). And a forthcoming paper (Koenig et al. forthcoming) takes a critical look at what sets ontologies and knowledge bases (and, in particular, Wikidata) apart from other forms of knowledge technology and argues that a specific kind of skillset and mindset, 'ontology literacy', is required to maintain true knowledge agency in an information context mediated by ontologies.

## References


Allison-Cassin, Stacy. 2022. "Libraries, Linked Data, and Decolonization." Keynote presented at the SWIB22, Berlin, November 28. https://swib.org/swib22/programme.html#libraries-linked-data-and-decolonization-keynote. Recording at: https://youtu.be/cJxfZSv4xEI.

Brückner, Sebastian, Florian Lemmerich, and Markus Strohmaier. 2021. "Inferring Sociodemographic Attributes of Wikipedia Editors: State-of-the-Art and Implications for Editor Privacy." In *Companion Proceedings of the Web Conference 2021 (WWW '21)*, 616–22. ACM. DOI 10.1145/3442442.3452350. Archived at

University of Mannheim MADOC: https://madoc.bib.uni-mannheim.de/61550/1/3442442.3452350.pdf.

Koenig, Loren. 2021. “Understanding Knowledge-Representation Inequities in Terms of Re-de-Contextualization.” Presented at the WikidataCon 2021, October 30. https://pretalx.com/wdcon21/talk/VQA9NQ/.

Koenig, Loren, Jennifer Stubbs, Alexandra Provo, and Megan Wacha. forthcoming. “Problematizing Metadata as Data. Ontology Literacy and Insight from the Wikiverse.”

Maitreyi, Maari. 2023. “Wikidata: Why We Contribute to the Robot Epistemology.” “*Whose Knowledge?” Blog* (blog). July 18, 2023. https://whoseknowledge.org/wikidata-robot-epistemology/.

Müller-Birn, Claudia, Benjamin Karran, Janette Lehmann, and Markus Luczak-Rösch. 2015. “Peer-Production System or Collaborative Ontology Engineering Effort: What Is Wikidata?” In proceedings of OpenSym ’15: The 11th International Symposium on Open Collaboration, 1–10. San Francisco California: ACM. DOI 10.1145/2788993.2789836. Archived at OpenSym website: https://www.opensym.org/os2015/proceedings-files/p501-mueller-birn.pdf.

Piscopo, Alessandro, and Elena Simperl. 2018. “Who Models the World? Collaborative Ontology Creation and User Roles in Wikidata.” *Proceedings of the ACM on Human-Computer Interaction* 2 (CSCW): 1–18. DOI 10.1145/3274410. Preprint at: https://eprints.soton.ac.uk/423194/1/cscw_text_review.pdf.

Provo, Alexandra, Megan Wacha, Loren Koenig, and Jennifer Stubbs. 2021. “Politics and Pedagogy of Wikidata in Libraries.” Roundtable presented at the ACRL 2021, online, April 15. https://airtable.com/app4udQN6VrZTUlQZ/shrawugUnnKE76tRg/tblIxIv5xO7lO4kTB/viwakeCGhN2PdU0CC/recVcTuQfH3BFhHVI. Etherpad notes at https://w.wiki/399c.

Sarasua, Cristina, Alessandro Checco, Gianluca Demartini, Djellel Difallah, Michael Feldman, and Lydia Pintscher. 2019. “The Evolution of Power and Standard Wikidata Editors: Comparing Editing Behavior over Time to Predict Lifespan and Volume of Edits.” *Computer Supported Cooperative Work (CSCW)* 28 (5): 843–82. DOI 10.1007/s10606-018-9344-y. Preprint at: https://eprints.whiterose.ac.uk/140352/1/evolution-wikidata-editors.pdf.

Simperl, Elena. 2018. “Loops of Humans and Bots in Wikidata.” In *Companion Proceedings* (proceedings of The Web Conference 2018 (WWW ’18), Lyon, France), 1107–1107. ACM Press. DOI 10.1145/3184558.3191552. Presentation slides available at https://www.slideshare.net/elenasimperl/loops-of-humans-and-bots-in-wikidata.

Zhang, Charles Chuankai, Mo Houtti, C. Estelle Smith, Ruoyan Kong, and Loren Terveen. 2022. “Working for the Invisible Machines or Pumping Information into an Empty Void? An Exploration of Wikidata Contributors’ Motivations.” *Proceedings of the ACM on Human-Computer Interaction* 6 (CSCW1): 1–21. DOI 10.1145/3512982.