Wikidata, historical lives, and the datafied ideologies of our present

James Baker (Southampton) and Ammandeep K. Mahal (Southampton)

"It might have been otherwise"
Susan Star 'Power, Technology, and the Phenomenology of Conventions'[1]

The rise of Wikidata represents a quiet revolution in knowledge infrastructure. Wikidata is a knowledge base, a source of linked open data on people, places, things, and concepts, designed to be read and edited by both people and machines. Wikidata is a central and language agnostic node within our contemporary knowledge infrastructure ecosystem. It is a federator of identifiers for everything from people and films to library classifications and mathematical software. It is a source of knowledge that is amplified by web content and digital assistants. It is a platform for knowledge production and for the machine-readable communication of research results.[2] And it is a model for the creation of new knowledge graphs, with Wikibase as-a-service underpinning platforms for exploring the lives of enslaved individuals,[3] OpenStreetMap metadata, and the work of various libraries and research institutes.

This long paper – based on research in its final stages – explores Wikidata as an infrastructure, what that infrastructure produces, and the implications of its centrality within our knowledge infrastructures and systems, as well as wider society. We are guided in our work by approaches from critical digital humanities, software studies, and – in particular – scholars of infrastructure and critical data studies whose research has collectively imagined these datafied knowledge infrastructures as embodiments of (not necessarily harmonious) community conventions and standards, and – drawing on the likes of Mary Douglas and Michel Foucault[4] – has analysed

[1] Susan Leigh Star, "Power, Technology and the Phenomenology of Conventions: On Being Allergic to Onions," *The Sociological Review* 38:1 (1990).
[2] See for example, Adriano Rutz et al., "The LOTUS Initiative for Open Knowledge Management in Natural Products Research," *ELife* 11 (2022).
[3] enslaved.org
[4] Mary Douglas, *Purity and Danger: An Analysis of Concepts of Pollution and Taboo* (1966); Michel Foucault, *The Archaeology of Knowledge* (1969).

DH Benelux, 31 May – 2 June, Brussels, Belgium

the classificatory logics they deploy as the products of a power and labour that is always relational and consequential.[5]

As one of the Wikimedia Foundation's family of sites Wikidata is maintained by community labour, specifically the labour of volunteers open collaborating using a wiki-based editing system. This community is particular and far from uniform. Previous studies seeking to understand who produces knowledge on Wikidata have focused on community structure, the relative heterogeneity of 'leaders', 'contributors', and bots, and the relationship between experienced editors and the ontological quality of the edits they make.[6] Less attention has been paid to the demographic profile and geographical spread of the Wikidata community. However, the under-representation of minoritised communities and demographics within the wider Wikimedia labour ecosystem is systemic, enduring, and well-documented, [7] with roughly one-tenth of Wikipedia editors identifying as a woman or non-binary and 20% of information in the encyclopaedia produced on or by people from the Global South.[8] This, coupled with knowledge that the infrastructures of the internet do not look or sound like most people in the world,[9] and a wealth of evidence on how classification, datafication, and machine processing deepens the injustices experienced by minoritised communities – especially so when that labour is undertaken without sufficient representation from those communities[10] – makes exploring the conventions, standards, and relational work that shape Wikidata as a classificatory infrastructure a vital task. It is also an imposing task. Wikidata contains close to

[5] Geoffrey C. Bowker and Susan Leigh Star, *Sorting Things out: Classification and Its Consequences* (2000); Daniela Agostinho, "Archival Encounters: Rethinking Access and Care in Digital Colonial Archives," *Archival Science* 19:2 (2019); Nanna Bonde Thylstrup et al., *Uncertain Archives Critical Keywords for Big Data* (2021).
[6] Alessandro Piscopo, "Structuring the World's Knowledge: Socio-Technical Processes and Data Quality in Wikidata" (PhD, 2019); Alessandro Piscopo and Elena Simperl, "Who Models the World?: Collaborative Ontology Creation and User Roles in Wikidata," *Proceedings of the ACM on Human-Computer Interaction* 2, no. CSCW (2018).
[7] Anasuya Sengupta and Siko Bouterse, "Research – Whose Knowledge?," 2017, whoseknowledge.org/why/; Mark Graham, Ralph K. Straumann, and Bernie Hogan, "Digital Divisions of Labor and Informational Magnetism: Mapping Participation in Wikipedia," *Annals of the Association of American Geographers* 105:6 (2015); Francesca Tripodi, "Ms. Categorized: Gender, Notability, and Inequality on Wikipedia," *New Media & Society* (2021).
[8] Az Causevic et al., "Centering Knowledge from the Margins: Our Embodied Practices of Epistemic Resistance and Revolution," *International Feminist Journal of Politics* 22:1 (2020); Anasuya Sengupta, "Decolonising Wikidata: Why Does Knowledge Justice Matter for Structured Data?", WikidataCon 2021 (2021), https://www.youtube.com/watch?v=wn2BrQomvFU.
[9] Sengupta, "Decolonising Wikidata: Why Does Knowledge Justice Matter for Structured Data?"
[10] Catherine D'Ignazio and Lauren F. Klein, *Data Feminism* (2020); Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (2018); Emily M Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" (2021); Anita Lavorgna and Pamela Ugwudike, "The Datafication Revolution in Criminal Justice: An Empirical Exploration of Frames Portraying Data-Driven Technologies for Crime Prevention and Control," *Big Data & Society* 8:2 (2021).

100 million items (the subjects and objects in any given triple), of which over 10,000 are properties (the predicates that bind those triples), 1.7 billion page edits, and close to 12 billion words of commentary and discussion.

Given this scale, rather than attempt to read Wikidata at scale, an approach that is common in the HCI and network studies literature, we employ of a narrower frame, a proverbial wedge that seeks to tease out the ideologies Wikidata has adopted, takes for granted, and is likely to continue to reproduce. As Willaert and Roumans show, narrow frames can provide highly elucidatory ways into complex knowledge systems.[11] Our frame is an in-development knowledge base called Beyond Notability that seeks to document women's work in archaeology, history, and heritage between 1870 and 1950, and that is the product of inclusive, reflexive, and positionally-inscribed cataloguing labour.[12] The ontology for Beyond Notability was designed alongside Wikidata, with Wikidata items and properties as a source both of alignment and tension. This paper then emerges from our deviations from Wikidata, from our active troubling of its assumptions and conventions, and from the spaces and perspectives that enabled us to explore the positionality of Wikidata, and in turn the concepts, people, and worldviews that are bent to fit its data model.

We proceed in three parts. First, we examine the presentness of Wikidata statements. Focusing on Wikidata properties – those verbal binds in subject-predicate-object triples – we reflect on how their presentness shapes the representation of historical relationships. We argue that having developed a data model around 'significant' events the Wikidata data model creates limited space for the rich, subtle, and granular conceptual changes required when representing complex historical phenomena. Second, we examine the ways in which Wikidata's approach to ascriptions of ethnicity and citizenship contribute to both a normalisation of unmarked whiteness and nationalist geopolitics in the presentation and use of Wikidata biographies. And third, we probe Wikidata's – not uncontroversial – collapsing of gender and sex, its normalisation of gender ascription, and its inattention to both how gender varies over time and between places.

---

[11] Tom Willaert and Guido Roumans, "Nitpicking Online Knowledge Representations of Governmental Leadership. The Case of Belgian Prime Ministers in Wikipedia and Wikidata.," *LIBER Quarterly* 30:1 (2020).
[12] Alexandra Ortolja-Baird and Julianne Nyhan, 'Encoding the Haunting of an Object Catalogue: On the Potential of Digital Technologies to Perpetuate or Subvert the Silence and Bias of the Early-Modern Archive'. *Digital Scholarship in the Humanities* (2021).

Throughout, we are guided by histories of "professional" women in twentieth-century Britain. Kate Hill has examined how women's work was constructed and imagined in museums.[13] Bonnie Smith has articulated the lack of job opportunities for women in the academic humanities.[14] Helen Glew, Helen McCarthy, and Claire Langhamer have all drawn attention to the societal constraints that shaped professional women's career development and authority in the workplace.[15] McCarthy in particular, has illuminated how in the second half of our period, an significant attitudinal shift took place that reconfigured the landscape of paid work for women and its relationship to traditional roles of motherhood and wifely household management.[16] It is shifts of this nature that, we argue, Wikidata fails to capture, and in foregrounding these tensions between particular historical phenomena and classificatory logics, our work stresses the value of using practice-based ontology development to reflexively critique, investigate, and probe knowledge infrastructures – on for data-driven research – in the age of platform web technologies.

[13] Kate Hill, *Women and Museums 1850-1914: Modernity and the Gendering of Knowledge* (2016).
[14] Bonnie G. Smith, *The Gender of History: Men, Women, and Historical Practice* (1998).
[15] Helen Glew, *Gender, Rhetoric and Regulation: Women's Work in the Civil Service and the London County Council, 1900-55* (2016); Helen McCarthy, *Double Lives: A History of Working Motherhood in Modern Britain* (2021); Claire Langhamer, "Feelings, Women and Work in the Long 1950s," *Women's History Review* 26:1 (2017).
[16] McCarthy, *Double Lives*, 136, 196.