Communication, Culture and Critique, 2024, **00**, 1–8
https://doi.org/10.1093/ccc/tcae029
Original Article

What does it mean to be queer in Wikidata? Practices of gender representation within a transnational online community

Beatrice Melis,[1] Chiara Paolini,[2] Marta Fioravanti,[3] Daniele Metilli [4,*]

[1]Department of Computer Science, University of Pisa, Pisa, Italy & Gran Sasso Science Institute, L'Aquila, Italy
[2]Department of Linguistics, KU Leuven, Leuven, Belgium
[3]oio.studio, London, UK
[4]Department of Information Studies, University College London, London, UK
*Corresponding author: Daniele Metilli. Email: d.metilli@ucl.ac.uk

Abstract

The continuing digitization and datafication that our society is undergoing are having a significant impact on our daily lives, giving rise to new possibilities but also entailing significant risks for people who are discriminated against or marginalized. Queer communities are particularly affected by these processes; therefore, it is crucially relevant to research transnational digital projects that involve them. In the Wikidata Gender Diversity (WiGeDi) project, we are looking at practices of gender representation in the Wikidata knowledge base, a collaborative online project managed by a worldwide community. Working from the idea that gender is a complex social construct, we investigate how the Wikidata community has approached the complex issue of modeling and populating gender data, progressing from a very narrow interpretation of gender as a binary to a representation that is more inclusive of a multiplicity of gender identities.

Keywords: Wikidata, gender diversity, online communities, critical data studies, queer studies.

The continuing digitization and datafication that our society is undergoing are having a significant impact on our daily lives (Kitchin, 2021; Pink et al., 2017). Every day, biographical data concerning our identities are searched, generated, shared, reused, and analyzed for different purposes. This technological turn has given rise to new possibilities for communication across national, societal, and cultural boundaries, but at the same time, it entails significant risks for people who are discriminated against or marginalized (Buolamwini & Gebru, 2018; Noble, 2018).

Many digital projects show a lack of care in the way data is handled (Puig de la Bellacasa, 2017), even when dealing with sensitive data such as gender identity and sexuality. Queer people, who are often underrepresented in popular online communities, are particularly impacted by these processes (Pain, 2022) and subject to exclusionary and hateful practices (Nash & Browne, 2020; Righetti, 2021). In this context, it is crucially relevant to research transnational digital projects that involve queer communities and look at the ethical decisions that are often made behind the scenes without careful evaluation by the communities, which may lead to unfairness, abuse, or erasure.

The Wikidata platform demonstrates in a concrete way the importance of careful data practices within a transnational digital community. Wikidata is a collaborative knowledge base funded and hosted by the Wikimedia Foundation and edited by a wide, international community of users (Vrandečić & Krötzsch, 2014). Wikidata was built as a central hub for other Wikimedia projects—such as Wikipedia, Wikimedia Commons, and Wikisource—allowing them to share and reuse structured data.

In the Wikidata Gender Diversity project (WiGeDi, https://wigedi.com), funded through the Wikimedia Research Fund and running from September 2022 to July 2024, we have looked at how the Wikidata community has approached the complex issue of modeling and populating gender data, focusing on trans, non-binary, and other gender-diverse identities (Metilli & Paolini, 2023). Working from the idea that gender is a complex social construct (Butler, 2004), we have studied gender representation in Wikidata through the lens of critical data studies (Iliadis & Russo, 2016; Kitchin & Lauriault, 2018). We have adopted a transfeminist, intersectional, and queer perspective (Butler, 1990; D'Ignazio & Klein, 2020; Lorde, 1984; Serano, 2007; Wittig, 1992). We reject binary, cis-heteronormative, and essentialist views of gender, recognizing the wide variety of gender identities and expressions that have existed throughout human history (DeVun, 2021; Herdt, 1994).

One of our main research questions, and the core focus of this article, concerns the work done by the Wikidata community, through the reconstruction of the historical progress that has seen the community incur several missteps and lack of care, but also try to remedy them through practices of data repair (Ramakrishnan et al., 2021) and changes in policies and technical infrastructures. Previous studies about gender in the Wikidata knowledge base do exist, but they have an approach that differs from ours, as they mostly focused on research into the *gender gap*, understood as the disparity in the representation of women compared to men. Klein et al. (2016) developed a biographical database integrating gender-related information across temporal, spatial, and occupational dimensions. Konieczny and Klein (2018) developed

Received: 1 December 2023. **Revised:** 25 June 2024. **Accepted:** 5 July 2024

indicators to measure worldwide longitudinal gender inequality trends. Zhang and Terveen (2021) analyzed the gender gap to identify biases that affect women's participation. More recently, Weathington and Brubaker (2023) conducted an ethnographic study on how queerness is represented in Wikidata. Gender and sexuality have been analyzed more extensively in Wikipedia (e.g., Beytía & Wagner, 2022; Field et al., 2022; Miquel-Ribé et al., 2021). Further relevant research concerns Wikidata's quality according to various metrics (e.g., Shenoy et al., 2022); and studies about the user community (e.g., Kaffee & Simperl, 2018; Piscopo et al., 2017).

In the following section, we introduce Wikidata and its history. After reporting our methodology, we discuss the Wikidata user community and how it has approached the representation of queerness. Later, we analyze issues in Wikidata practices around gender diversity, looking at the model, population of gender data, and automation through bots. We then look at how the Wikidata community has implemented practices of data repair. Finally, we discuss what insights can be gathered from the results of our project and provide our conclusions.

Wikidata

Wikidata is unique among large knowledge bases because it is not managed by a corporation, but rather, it is self-governed by a community of users (Piscopo et al., 2017). Similarly to Wikipedia and other projects hosted by the Wikimedia Foundation, Wikidata is run by volunteer users who work as editors to build and maintain a free, open-source database of structured data. Anyone can contribute to Wikidata, and through a virtuous cycle of collaboration, everyone benefits from the contributions of others and the quality of the project improves over time.

One of the key elements that facilitates collaboration in wiki platforms is their open-source and non-proprietary nature (Seiler & Aaltonen, 2014). This allows for a level of transparency and accessibility that encourages people to get involved. These collaborative and dynamic platforms need practical configurations to agree and ensure respect between users in relation to published content, and this is why projects such as Wikipedia and Wikidata have adopted a consensus-building method. A consensus is achieved through community discussion and voting on proposals to add or modify data and information. When a proposal is made, other users review the proposal and provide feedback. If there is general agreement among the community, the proposal is approved, and the procedure is completed. In addition, Wikidata has been designed as an inherently multilingual project (Kaffee et al., 2017), which means that it is not divided into subprojects covering specific languages; instead, it is a unified platform that connects its community in a language-independent way (Kaffee & Simperl, 2018). Each time the community needs to decide how to approach the representation of knowledge within Wikidata, it must come together and overcome language, geographical, and cultural barriers to agree on a set of guidelines to structure data.

This process may work well when modeling concepts on which there is universal understanding and agreement, but queerness is particularly difficult to model because it inherently resists definition and defies classification. Queerness is outside, it is beyond, it does not fit. In our view, there is no data model that can faithfully represent the wide variety, multiplicity, and fluidity of queer identities, because these identities are socially constructed and their definitions are not universally agreed (Butler, 2004). This is especially true when considering the variation in the understanding of queerness over history, the evolution of language, and how anyone's perception of their own identity can evolve over time.

Nonetheless, Wikidata users have been attempting to achieve this goal by creating a common data model, or *ontology*, which has been developed over the years by the community. In computer science, an ontology is "a formal, explicit specification of a shared conceptualization" of a certain domain of reality (Guarino et al., 2009). In Wikidata, this is realized through the definition of a taxonomy of classes, and properties expressing relations among them. One of Wikidata's aims is to represent biographical knowledge about people, including sensitive data such as their gender. Our research involves an in-depth analysis of how gender identities are conceptualized and represented in the Wikidata ontology.

Methods

To discuss the methodology we have adopted, we first need to acknowledge our own positionality. We are queer White Italian early-career researchers based in the Global North (Italy, the United Kingdom, and Belgium). The primary authors, Beatrice and Daniele, have previous experience as users of Wikidata and related platforms, and Daniele is a former administrator who was active in the early years of Wikidata. Our previous experiences as queer users on these platforms have significantly shaped our project.

To fully understand how gender is represented in Wikidata, our project has adopted a multi-faceted approach that looks at gender from three complementary perspectives: *model*, *data*, and *community*. This choice has been influenced by our backgrounds: Daniele has a background in knowledge representation, Beatrice and Marta have experience in data analysis and visualization, and Chiara is a researcher in linguistics. While other possible methods exist (e.g., community surveys), our perspectives cover the main research questions that can be asked about gender representation on the platform, namely, *how* gender is represented, *who/what* is represented, and *why*.

To begin with, our project has investigated how Wikidata currently represents gender. We attempt to understand the extent to which this representation is inclusive of queer identities, and how it supports the representation of a wider spectrum of identities by looking at how the Wikidata ontology model has evolved over time. Beyond looking at the model, we have also analyzed the data stored in the knowledge base in a quantitative way, to gather insights and identify possible gaps related to gender diversity. Specifically, we have examined the distribution of gender diverse identities across space, time, and other dimensions. Our findings show a clear underrepresentation of these populations in Wikidata compared to their prevalence of these populations in society (Metilli & Paolini, 2023). We have also looked 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, etc. Finally, we have been 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 platform such as Wikidata. Therefore, we analyze user discussions about the topic of gender identities through both qualitative and quantitative methods.

These techniques have been applied to a corpus of user discussions in English on gender-related topics. In our initial analysis, we started from a set of 1,829 user discussions from 2012 to 2022. These discussions were selected because they contained relevant terms from a set of 79 terms related to gender, sex, and queerness. The discussions were then filtered through an annotation interface of our own development,[1] resulting in a corpus of 604 relevant discussions. The choice of considering discussions in English was driven by the fact that this is the de facto primary language used on the platform when different communities need to interact with each other. Different language communities use language-specific "talk pages" to hold discussions, but when users who speak different languages need to talk to each other, they usually communicate in English.

For this reason, we decided to focus primarily on English, though we acknowledge the limitations of our current approach and plan to extend the analysis to non-English discussions in the future. Nonetheless, the analyses we conducted have also considered a wider range of languages outside of user discussions—for example, in the analysis of the gender model, we conducted a qualitative analysis of labels and descriptions of classes and properties from the Wikidata ontology (Melis, 2023) that informed this study. The results presented in this article are derived from the outcomes of the three methodological perspectives listed above, i.e., model, data, and community. By adopting this multi-faceted approach, the project has gained a deeper understanding of gender diversity in Wikidata.

Queerness in a transnational community

Since the launch of Wikidata in 2012, users have always been the driving force behind the success of the platform, but in the early years, the community was small and not very diverse. Based on the results of the 2021 Community Survey (Wikimedia Deutschland, 2021)—the only large-scale survey that has been made on the topic—the early community was composed mostly of men from the Global North. While the situation has improved over time, as of 2021, about 75% of users were still male and about 76% were still based in the Global North. Meanwhile, the community has grown significantly, reaching 12,000 active editors as of September 2023 (Wikimedia Foundation, 2013).

In general, the composition of the Wikidata community plays a significant role in shaping the platform's approach to the representation of queer identities, since it may not fully reflect broader societal demographics. It is beneficial for any collaborative platform to have a diverse community, with a plurality of perspectives and cultural backgrounds. In the early years of Wikidata, decisions about gender modeling were often taken by a small group of users, while the perspectives of marginalized identities and cultures were underrepresented. This lack of representation led to oversights in the representation of queerness, including gender diversity.

Gender is a highly complex topic that is very difficult to model, and unfortunately, not all users within the early Wikidata community possessed a deep understanding of this complexity. Some contributors approached gender modeling with a simplistic and narrow perspective, failing to account for the cultural, linguistic, and historical variations in gender identity and expression. This lack of awareness resulted in a lack of care about gender data, leading to misrepresentations, stereotypes, and inaccuracies.

Furthermore, in the early years of Wikidata, there were no clear rules or policies around biographical data, and when a new proposal was made, only a few users participated in discussions. When looking at early discussions (e.g., Wikidata, 2013), we see that this group of users lacked a clear understanding of the complexity of gender. Being influenced by cisheteronormativity, they tended to view gender as a binary and believed that every person described in the knowledge base should necessarily be classified according to their gender. Indeed, in the early user discussions there was no mention of the existence of trans, non-binary, and other gender diverse identities.

In this regard, it is important to note that Wikidata contributors have diverse backgrounds, and their priorities may vary widely. While some may insist on creating an accurate gender model and compiling data in a careful way, others may prioritize different areas of knowledge or have distinct concepts of *care*. For example, in the early discussions, users seemed to care mostly about increasing completeness metrics, such as the number of people described in the knowledge base, or the number of people with an assigned gender, as explained below in the section on Modeling Conflicts.

This divergence in priorities resulted in an inconsistent commitment to ensuring accurate gender representation. Efforts to address this issue should involve fostering a sense of shared responsibility and raising awareness about the significance of comprehensive gender data—but this represents a complex challenge for online collaborative platforms.

Moreover, given the current anti-queer and anti-trans climate (Pearce et al., 2020), these platforms may become a battleground between inclusive and narrow views of gender, influenced by both the users' culture and their personal views about the topic.

Another important aspect to consider is the transnational and multilingual nature of Wikidata, whose community is loosely divided into subgroups based on language, which often clash when discussing the representation of gender. While topic-based user groups exist, most of them are not centered on users' identities, and it is often challenging for users to find common ground. Conflicts in the community affect the model itself, descriptions of individuals, and user discussions. For example, trans people are discriminated against in the model, where their gender identities have been represented separately from cisgender identities, and for a long time, trans women were not included among women and trans men were not included among men. Moreover, descriptions of trans people in Wikidata are affected by frequent "edit wars" among users with different viewpoints, especially when someone makes their identity public or announces a gender transition. Proposals to make the model more inclusive face resistance, require long and extensive discussions, and are sometimes rejected altogether (e.g., Wikidata, 2017).

A further significant issue is that Wikidata users have relied heavily on bots and automated tools to populate gender data, as we will see in the section on The Rise of Bots. This data has historically been imported without much human oversight from Wikipedia and other sources such as the Virtual

International Authority File. In some cases, users have relied on personal names (or occasionally pronouns) to infer someone's gender, but this is prone to significant errors, because the same personal name can be considered masculine, feminine, or neutral depending on the culture or language.

These practices show that, especially in the early years of Wikidata, gender data was handled with insufficient care, and while the community has since achieved a greater awareness of the issue, it is still far from being fully inclusive towards queer people. It is interesting to note that different forms of queerness are treated differently in Wikidata, as data about sexual orientation follows significantly different rules compared to gender. Indeed, the platform contains warnings stating that sexual orientation data should not be added except when individuals have "stated it themselves, unambiguously, or it has been widely agreed upon by historians after their death" (Wikidata, 2024). No such rule exists for gender data, where historically, the community assumed that every human should be assigned a gender value.

In our project we have chosen to look primarily at gender identities, adopting the three perspectives of model, data, and community. In the next section, we report our findings about the Wikidata gender model.

Modeling conflicts

As Wikidata opened to the public in October 2012, a series of important events happened in subsequent months, shaping the foundation of how gender would be conceptualized in the knowledge base. Before going into the issues with the current representation of gender in Wikidata, we need to discuss the basics of how gender is modeled in its ontology. The property that represents gender in Wikidata is P21, currently labeled "sex or gender." The purpose of this property is to represent the sex or gender of people and any other organisms (e.g., animals or plants), both real and fictional. In this article, we focus exclusively on people.

A "gender" property was initially proposed by a Wikidata user on February 2, 2013 (Wikidata, 2013). At the time, the platform averaged about one hundred active users (Wikimedia Foundation, 2013), yet only five people participated in the discussion about the creation of the property. Different points of view emerged, with considerations influenced by a binary perspective centered on sex assigned at birth (male, female, intersex) instead of gender identity. After just two days of deliberation, the proposal was approved on February 4, 2013. The property was initially labeled "gender" in some languages (e.g., English and French) and "sex" in others (e.g., German and Italian). After a long discussion, the English label of P21 was changed to "sex (or gender)" in December 2013—and the following month the parentheses were removed. The property has since been labeled "sex or gender."

The first inherent problem with the P21 property, and consequently with Wikidata's gender model, is closely related to this property label, which has been the subject of discussions since its creation. The label "sex or gender" is ambiguous because it conflates multiple concepts, including sex assigned at birth, gender identity, and gender expression. In this regard, it is interesting to note that more than one value can be associated with this property in Wikidata, for example, a person may be assigned "trans woman" and "intersex" values at the same time. While this feature enables a more nuanced representation, it may also lead to ambiguity, as there is no clear distinction between values relating to sex assigned at birth and those relating to gender identity or expression. This may contribute to confusion and possible misinterpretation of gender data.

When editing biographical data, Wikidata users must follow certain policies and guidelines approved by the community. These include, for example, the policy on data about living people (Wikidata, 2023a) and others around notability and verifiability of data. When it comes to P21, the community has also imposed significant constraints on the allowed values of the property—these include instances of *gender identity*, instances of *sex*, but also other values which are not instances of either of these two concepts (e.g., "third gender"), as well as the *unknown* value (Metilli & Paolini, 2023). However, it is important to note that these constraints do not prevent the users from setting any value of their choice and are simply used to check for errors after-the-fact.

P21 is not the only Wikidata property that expresses information about people's gender. Indeed, there are other properties that may communicate gender, and especially those used to express family relations. Due to the reliance on a cis-heteronormative binary view of gender, the community encountered significant challenges with the incorporation of gender-neutral properties within the modeling of family relations. In particular, the binary properties "mother/father," "brother/sister," "uncle/aunt," prevented the recognition of non-binary individuals. Although concerns about these gendered properties emerged early in the platform's history, effecting meaningful change was a long and difficult process. An important event was the community's decision to initiate a transition in 2013 by replacing the conventional designations of "uncle/aunt" (and other non-immediate family relations) with the gender-neutral property "relative." Subsequently, in 2017, the properties "brother/sister" and "stepfather/stepmother" were similarly replaced, respectively, with the neutral properties "sibling" and "stepparent." However, the replacement of "brother/sister" was accompanied by contentious discussions, as a segment of the Wikidata community resisted the change due to linguistic issues, because certain languages (e.g., Italian) lack a direct equivalent for the term "sibling." Problems of this nature take different forms and extent depending on the language used. Even when the Wikidata community adopted gender neutral terms since the creation of the property, as in the case of "spouse", there are languages in which a neutral translation of the label is not possible. For example, while in Italian a neutral term "coniuge," is used as the label of the property, in German the grammatical gender is made explicit through the label "Ehepartner(in)," representing the masculine term "Ehepartner" and the feminine one "Ehepartnerin."

The most important case of binary family relation properties that is still unresolved is that of "mother/father." Since English has a neutral term "parent," multiple proposals have been made over the years advocating for the substitution of these gendered designations with the gender-neutral term. Despite multiple attempts, these proposals encountered hurdles and did not manage to reach the widespread consensus necessary for approval. The most recent development is the creation of a "parent" property, which however is only applied to people whose gender is non-binary or unknown, thereby perpetuating exclusion. This process of reevaluating family-related properties highlights the continuing attention that the Wikidata community pays to these issues, despite

significant challenges along the way, principally due to linguistic nuances and the difficulty in reaching the broad consensus required to achieve change.

Population of gender data

The process of populating gender information on Wikidata involves collecting data from several sources. This section discusses the different methods through which gender data is entered into Wikidata and highlights the complexities associated with each approach. One of the main methods of populating gender information on Wikidata is manual annotation by users. This process involves individual contributors reviewing and verifying the gender data of items. Although manual annotation allows for a certain degree of accuracy, it can also be unintentionally subjective—and consequently prone to errors, as the interpretation of gender may vary according to cultural, linguistic, ideological, and personal biases. Furthermore, discrepancies may occur when multiple users annotate the same entry, potentially leading to inconsistent data.

Wikidata users have relied extensively on the sister project Wikipedia as a source of gender data. Unfortunately, Wikipedia is not a reliable source, as it is itself user-generated and prone to errors. Indeed, Wikipedia may contain inaccuracies or lack proper references for a person's gender, and consequently, using Wikipedia as a direct source for gender data may perpetuate misinformation.

External databases and authority files constitute an additional source of gender information for Wikidata, but their reliability varies greatly, hence this approach can also present significant challenges. Some databases may have well-established and accurately compiled gender data, while others may lack proper verification—or worse, may reflect biases in the original sources. To make things more complicated, some databases may be reliable when it comes to certain data, e.g., dates of birth, but not respect people's self-identification when it comes to gender.

The most problematic approach that has been adopted on a large scale by the Wikidata community is the population of gender data based on personal names, which is an extremely problematic procedure because personal names can be associated to different genders across cultures and languages, leading to frequent misclassification and a reinforcement of gender stereotypes. Another method that is highly questionable—though fortunately it has not been applied on a large scale on Wikidata—is the reliance on pronouns to determine gender. The use of pronouns can vary according to linguistic and cultural factors and is not necessarily related to a person's gender identity or sex assigned at birth. Furthermore, limited pronoun options in certain languages make this approach very unreliable.

Evidently, the population of gender information within Wikidata is a complex process, and since it concerns sensitive data about people, it requires careful approaches, as well as continuous supervision and verification. The methods mentioned above present several challenges, ranging from cultural and linguistic variations to potential biases and inaccuracies. This issue is compounded by the reliance on massive data imports without supervision, which shows a lack of care for gender data. Indeed, manual editing can reduce the chance of causing harm, which is one of the goals of Wikidata's policy on data about living people (Wikidata, 2023a). In the next section, we will look at the historical and current use of bots and other automated tools for the population of gender data.

The rise of bots

In Wikidata, bots play a significant role, acting as automated tools to perform repetitive, complex, or extensive tasks. They help maintain and improve the completeness and consistency of the data, since they can be used to make edits on a large scale without direct human intervention. In addition, bots facilitate more efficient collaboration by handling trivial edits and allowing human editors to focus on more important tasks. However, when bots are used in a careless and unsupervised way, as it happened with gender data in Wikidata, they may cause significant inaccuracies that are difficult to identify and repair after-the-fact.

On February 6, 2013, just two days after the establishment of Wikidata's gender property (P21), users started to seek permission to employ bots for the purpose of adding gender data. This proposal aimed to use information derived from Wikipedia categories, with the rationale that certain categories, such as "Women physicists," could be safely assumed to be associated with a specific gender. This proposal gained approval in the span of two weeks, marking the commencement of automated gender data population.

Subsequently, this automated process became a routine practice, often relying on personal names to guess gender identity. Over two years, at least seven distinct bots, developed by different users, were deployed to augment Wikidata with gender information. By June 2015, gender data completeness was reported to have reached 93.8%. The systematic process of gender data addition via bots introduced significant inaccuracies in Wikidata, necessitating manual repair by users; in many cases, however, the items remained incorrect for a long time before a user noticed the error.

Nevertheless, these practices have persisted. In March 2023, a discussion unfolded regarding the practice of assigning gender information to individuals based on given names. This action had been carried out by a single user to address the under-representation of women in the knowledge base through an automated tool. However, concerns were raised about the accuracy and potential harm related to this approach. Users participating in the discussion debated the ethical implications of this procedure, emphasizing that gender identity does not necessarily correlate with given names. For example, Tibetan people with the word "Lama" in their name were misgendered because this honorific title was mistaken for the Arabic given name "Lama." The discussion highlighted the tension between equitable representation and responsible data practices, with some advocating manual editing to ensure accuracy and avoid harm. Ultimately, the user who had added the data agreed to revert the edits.

As we will see in the following section, multiple initiatives are working to address the historical and present issues around automated addition of gender data. However, due to the extensive scope of bot-driven activities, it is likely that many errors persist within the current version of Wikidata.

Practices of data repair

The acknowledgment of gaps and inaccuracies in gender representation within Wikidata has encouraged initiatives aimed

at "data repair." We view this concept as any attempt at addressing lack of care in data representation (in this case, gender modeling), reflecting Puig de la Bellacasa's (2017) work on *matters of care*, Zakharova and Jarke's (2024) definition of *care-ful data studies* ("a care lens allows to foreground how communities and their collective responses to datafication aim to shift power and care relations," p. 652), and Ramakrishnan et al.'s (2021) work on repair practices in technology infrastructures. In the following, we describe the data repair practices that we have identified in Wikidata.

Initiatives by individual users. Individual contributors form the backbone of Wikidata's efforts to repair gender data inaccuracies. These dedicated individuals work to ensure that gender data aligns with cultural, linguistic, and historical facets. Their efforts showcase the power of community-driven data repair within the context of a large collaborative ecosystem.

Initiatives by internal user groups. Internal user groups, such as WikiProject LGBT (Wikidata, 2023b), consist of volunteers who share a common interest in improving gender and LGBT-related content on Wikidata. Project members collaboratively address gender representation issues, correct misclassifications, and contribute contextual information to remedy content gaps.

Initiatives by users across different projects. Collaborative events, such as the Queering Wikipedia conference (Wikimedia Foundation, 2023), provide a channel for users across different projects to hold discussions, workshops, and collaborations centered around queerness, thereby leveraging collective expertise to repair data inaccuracies and incompleteness.

Initiatives by advocacy groups. Advocacy groups such as Art+Feminism (Art+Feminism, 2020) have emerged as a powerful force in remedying gender data gaps through edit-a-thons that add and correct information related to underrepresented gender identities. These groups actively engage in data repair practices and promote awareness of issues around diversity and inclusion.

As these practices continue to evolve and expand, they contribute to the ongoing transformation of Wikidata into a platform that more accurately reflects the complexity and diversity of gender and sexuality.

Discussion

WiGeDi has been the first research project to systematically investigate gender diversity in Wikidata with a specific focus on marginalized gender identities. In the current anti-queer and anti-trans climate (Pearce et al., 2020), Wikidata has become a focal point for contentious debates regarding gender representation. These conflicts are not only rooted in personal ideologies about gender but are also deeply intertwined with the cultural backgrounds of the platform's users (Metilli & Paolini, 2023).

In the previous sections, we discussed different types of conflicts within Wikidata's transnational and multilingual community when it debated gender representation in the knowledge base. These clashes can be attributed to different cultural norms, values, and perspectives on gender. Our findings shed light on the challenges and difficulties faced by queer communities within the vast data landscape of Wikidata.

The results of our project lead us to conclude that a shared representation of gender that is inclusive towards queer identities has not yet been fully achieved on Wikidata, although there has been a positive evolution over time. Throughout the project, we have often reflected on the meaning of the term "inclusive," which may suggest that there exists a majority normative group that needs to include some minorities, thus *othering* them. On the contrary, we believe that true inclusion will be achieved when multiple identities (also beyond gender) can positively coexist and thrive in collaborative online communities.

Previous studies have found that queer people are underrepresented and discriminated against in online communities (Pain, 2022; Nash & Browne, 2020; Righetti, 2021). This has been shown by Nafus (2012a) in open-source communities,[2] by Bivens (2017) on Facebook, and by Miquel-Ribé et al. (2021) on Wikipedia. Our results confirm some of these findings by identifying several cases of misrepresentation of gender identities that affect queer communities. In particular, the differences detected by Miquel-Ribé among Wikipedia language editions are reflected in the language-related conflicts that we have identified in Wikidata user discussions. Moreover, we have replicated some of the results of previous studies on gender in Wikidata, namely, the existence of a large gender gap previously identified by Konieczny and Klein (2018) and Zhang and Terveen (2021). In addition, we have measured the underrepresentation of marginalized gender identities (Metilli & Paolini, 2023) and we have independently observed the issues reported by Weathington and Brubaker (2023) around the difficulty of representing queer identities in Wikidata.

Another important aspect that had not been highlighted in previous studies is that in many cases, gender-related conflicts among Wikidata users stem from their language and/or their cultural understanding of gender due to the transnational and multilingual nature of the platform, as exemplified by the conflicts in the Wikidata community that we have highlighted in this article. Through an examination of historical missteps, underrepresentation, and data population issues, our study underscores the need for thoughtful and sensitive approaches to gender data in a transnational context.

Conclusion

The WiGeDi project presents an exploration of the interplay between transnational queer cultures and the evolving landscape of gender representation in the digital sphere, adopting as a case study the Wikidata knowledge base. The multifaceted methods employed in our research shed light on the path from initial limitations and inaccuracies in gender modeling to ongoing efforts toward greater inclusion of queer identities. Throughout a span of more than 10 years, the Wikidata community has worked across national borders to develop a shared understanding of gender. Our examination of the Wikidata model, data, and community provides a valuable case study that underscores the transformative potential of collaborative digital environments.

Through an examination of historical missteps and underrepresentation, our analysis highlights the need for thoughtful and culturally sensitive approaches to data management in a transnational context. While it is necessary to overcome cultural and linguistic barriers in a broad collaborative context, finding agreement among different viewpoints and experiences, this should not come at the expense of a representation that faithfully expresses the complexity of gender.

The results of our work underscore the importance of careful data practices in transnational digital projects to achieve

meaningful data justice (Taylor, 2017) for queer communities around the world. By uncovering the challenges faced by the Wikidata community and the subsequent practices of data repair, our research provides insights into how a digital platform that sits at the center of the web of data can shape narratives around gender. Ultimately, we hope that our research will help develop more inclusive data policies towards queer people in Wikidata and beyond.

Data availability

The data we relied upon is publicly available through Wikidata and can be downloaded using appropriate queries (see https://wigedi.com/data). The source code of our tools will be published on GitHub (https://github.com/wigedi) after related publications undergo peer review.

Funding

This study was supported with funding from the Wikimedia Research Fund, grant number G-RS-2204-08611.

Conflicts of interest: The authors report no conflicts of interest.

Notes

1. See Data Availability section for information about the publication of our source code.
2. See also the response by Dunbar-Hester and Coleman (2012) and the further rebuttal by Nafus (2012b).

References

Art+Feminism (2020). *About—Art+Feminism*. Retrieved June 25, 2024 from https://artandfeminism.org/about/

Beytía, P., & Wagner, C. (2022). Visibility layers: A framework for facing the complexity of the gender gap in Wikipedia content. *Internet Policy Review*, *11*(1). https://doi.org/10.14763/2022.1.1621

Bivens, R. (2017). The gender binary will not be deprogrammed: Ten years of coding gender on Facebook. *New Media & Society*, *19*(6), 880–898. https://doi.org/10.1177/1461444815621527

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler, & C. Wilson (Eds.), *Proceedings of machine learning research: Vol 81. Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (pp. 77–91). ML Research Press. https://proceedings.mlr.press/v81/buolamwini18a.html

Butler, J. (1990). *Gender trouble: Feminism and the subversion of identity*. Routledge. https://doi.org/10.4324/9780203824979

Butler, J. (2004). *Undoing gender*. Routledge. https://doi.org/10.4324/9780203499627

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT Press. https://doi.org/10.7551/mitpress/11805.001.0001

DeVun, L. (2021). *The shape of sex: Nonbinary gender from Genesis to the Renaissance*. Columbia University Press. https://doi.org/10.7312/devu19550

Dunbar-Hester, C., & Coleman, G. (2012, June 26). Engendering change? Gender advocacy in open source. *Culture Digitally*. http://culturedigitally.org/2012/06/engendering-change-gender-advocacy-in-open-source/

Field, A., Park, K. Z. L., & Tsvetkov, Y. (2022). Controlled analyses of social biases in Wikipedia bios. In F. Laforest, R. Troncy, E. Simperl, D. Agarwal, A. Gionis, I. Herman, & L. Médini (Eds.), *Proceedings of ACM Web Conference 2022* (pp. 2624–2635). Association for Computing Machinery. https://doi.org/10.1145/3485447.3512134

Guarino, N., Oberle, D., & Staab, S. (2009). What is an ontology? In: Staab, S. & Studer, R. (Eds.), *Handbook on Ontologies* (pp. 1–17). Springer. https://doi.org/10.1007/978-3-540-92673-3

Herdt, G. (1994). *Third sex, third gender*. Zone Books. https://doi.org/10.2307/j.ctv16t6n2p

Iliadis, A., & Russo, F. (2016). Critical data studies: An introduction. *Big Data & Society*, *3*(2), 205395171667423. https://doi.org/10.1177/2053951716674238

Kaffee, L. A., Piscopo, A., Vougiouklis, P., Simperl, E., Carr, L., & Pintscher, L. (2017). A glimpse into Babel: An analysis of multilinguality in Wikidata. In L. Morgan (Ed.), *OpenSym '17: Proceedings of the 13th International Symposium on Open Collaboration*. Association for Computing Machinery. https://doi.org/10.1145/3125433.3125465

Kaffee, L. A., & Simperl, E. (2018). Analysis of Editors' languages in Wikidata. In L. Morgan (Ed.), *OpenSym '17: Proceedings of the 13th International Symposium on Open Collaboration*. Association for Computing Machinery. https://doi.org/10.1145/3233391.3233965

Kitchin, R. (2021). *Data lives: How data are made and shape our world*. Bristol University Press.

Kitchin, R., & Lauriault, T. (2018). Toward critical data studies: Charting and unpacking data assemblages and their work. In J. Thatcher, J. Eckert, & A. Shears (Eds.), *Thinking big data in geography: New regimes, new research* (pp. 3–20). University of Nebraska Press.

Klein, M., Gupta, H., Rai, V., Konieczny, P., & Zhu, H. (2016). Monitoring the gender gap with Wikidata Human Gender Indicators. In A. I. Wasserman (Ed.), OpenSym '16: Proceedings of the 12th International Symposium on Open Collaboration. Association for Computing Machinery. https://doi.org/10.1145/2957792.2957798

Konieczny, P., & Klein, M. (2018). Gender gap through time and space: A journey through Wikipedia biographies via the Wikidata Human Gender Indicator. *New Media & Society*, *20*(12), 4608–4633. https://doi.org/10.1177/1461444818779080.

Lorde, A. (1984). *Sister outsider*. Crossing Press.

Melis, B. (2023). Wikidata Gender Diversity (WiGeDi): *Studying gender representation in Wikidata through the lens of data, model, and community* (Publication no. etd-09142023-125318) [Master's thesis, University of Pisa]. ETD. https://etd.adm.unipi.it/theses/available/etd-09142023-125318/

Metilli, D., & Paolini, C. (2023). Non-binary gender representation in Wikidata. In B. M. Watson, A. Provo, & K. Burlingame (Eds.), *Ethics in linked data* (pp. 221–264). Litwin Books.

Miquel-Ribé, M., Kaltenbrunner, A., & Keefer, J. M. (2021). Bridging LGBT+ content gaps across Wikipedia language editions. *The International Journal of Information, Diversity, & Inclusion (IJIDI)*, *5*(4), 90–131. https://doi.org/10.33137/ijidi.v5i4.37270

Nafus, D. (2012a). 'Patches don't have gender': What is not open in open source. *New Media & Society*, *14*(4), 669–683. https://doi.org/10.1177/1461444811422887

Nafus, D. (2012b, July 1). A response to 'Engendering Change?' *Culture Digitally*. http://culturedigitally.org/2012/07/a-response-to-engendering-change/

Nash, C. J., & Browne, K. (2020). *Heteroactivism: Resisting lesbian, gay, bisexual and trans rights and equalities*. Zed Books. https://doi.org/10.5040/9781350225503

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press. https://doi.org/10.2307/j.ctt1pwt9w5

Pain, P. (Ed.). (2022). *LGBTQ digital cultures: A global perspective*. Routledge. https://doi.org/10.4324/9781003196457

Pearce, R., Erikainen, S., & Vincent, B. (2020). TERF wars: An introduction. *The Sociological Review*, *68*(4), 677–698. https://doi.org/10.1177/0038026120934713

Pink, S., Sumartojo, S., Lupton, D., & Heyes La Bond, C. (2017). Mundane data: The routines, contingencies and accomplishments of digital living. *Big Data & Society*, *4*(1), 205395171770092. https://doi.org/10.1177/2053951717700924

Piscopo, A., Phethean, C., & Simperl, E. (2017). What makes a good collaborative knowledge graph: Group composition and quality in Wikidata. In G. L. Ciampaglia, A. Mashhadi, & T. Yasseri (Eds.), *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I* (pp. 305–322). Springer. https://doi.org/10.1007/978-3-319-67217-5_19

Puig de la Bellacasa, M. (2017). *Matters of care: Speculative ethics in more than human worlds*. University of Minnesota Press.

Ramakrishnan, K., O'Reilly, K., & Budds, J. (2021). The temporal fragility of infrastructure: Theorizing decay, maintenance, and repair. *Environment and Planning E: Nature and Space*, *4*(3), 674–695. https://doi.org/10.1177/2514848620979712

Righetti, N. (2021). The anti-gender debate on social media. A computational communication science analysis of networks, activism, and misinformation. *Comunicazione Politica*, *23*(2), 223–250. https://doi.org/10.3270/101610

Seiler, S., & Aaltonen, A. (2014). Wikipedia: The value of open content production. *Centre for Economic Policy Research*. https://cepr.org/voxeu/columns/wikipedia-value-open-content-production

Serano, J. (2007). *Whipping girl: A transsexual woman on sexism and the scapegoating of femininity*. Seal Press.

Shenoy, K., Ilievski, F., Garijo, D., Schwabe, D., & Szekely, P. (2022). A study of the quality of Wikidata. *Journal of Web Semantics*, *72*, 100679. https://doi.org/10.1016/j.websem.2021.100679

Taylor, L. (2017). What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society*, *4*(2), 205395171773633. https://doi.org/10.1177/2053951717736335

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

Weathington, K., & Brubaker, J. R. (2023). Queer identities, normative databases: Challenges to capturing queerness on Wikidata. *Proceedings of the ACM on Human-Computer Interaction*, *7* (CSCW1), 1–26. https://doi.org/10.1145/3579517

Wikidata (2013, February 3). *Property proposal archive 1*. Retrieved June 25 2024 from https://www.wikidata.org/wiki/Wikidata:Property_proposal/Archive/1#Gender_/_Geschlecht_/_Genre_(sexe)

Wikidata (2017, January 13). Requests for comment: Make family member properties gender neutral. Retrieved June 25 2024 from https://www.wikidata.org/w/index.php?title=Wikidata:Requests_for_comment/Make_family_member_properties_gender_neutral&oldid=428439241

Wikidata (2023a, July 27). *Living people*. Retrieved June 25 2024 from https://www.wikidata.org/w/index.php?title=Wikidata:Living_people&oldid=1941487793

Wikidata (2023b, July 29). *WikiProject LGBT*. Retrieved June 25 2024 from https://www.wikidata.org/w/index.php?title=Wikidata:WikiProject_LGBT&oldid=1942337671

Wikidata (2024, May 25). *Property P91*. Retrieved June 25, 2024, from https://www.wikidata.org/w/index.php?title=Property:P91&oldid=2163786174

Wikimedia Foundation (2023, July 26). *Queering Wikipedia*. Meta. Retrieved June 25 2024 from https://meta.wikimedia.org/w/index.php?title=Queering_Wikipedia&oldid=25350566

Wikimedia Foundation (2013). *Wikidata—Active* editors. Retrieved June 25, 2024, from https://stats.wikimedia.org/#/wikidata.org/contributing/active-editors/normal%7Cline%7Call%7C(page_type)~contentnon-content%7Cmonthly

Wikimedia Deutschland (2021). *Wikidata Community Survey 2021* (pp. 28–29). https://commons.wikimedia.org/w/index.php?title=File:Wikidata_Community_Survey_2021.pdf

Wittig, M. (1992). *The straight mind and other essays*. Beacon Press.

Zakharova, I., & Jarke, J. (2024). Care-ful data studies: Or, what do we see, when we look at datafied societies through the lens of care? *Information, Communication & Society*, *27*(4), 651–664. https://doi.org/10.1080/1369118X.2024.2316758

Zhang, C. C., & Terveen, L. (2021). Quantifying the gap: A case study of Wikidata gender disparities. In G. Robles, J. Arroyo, A. Barcomb, K. K. Chahal, S. Sowe, & X. Wang (Eds.), *OpenSym '21: Proceedings of the 17th International Symposium on Open Collaboration*. Association for Computing Machinery. https://doi.org/10.1145/3479986.3479992

Communication, Culture and Critique, 2024, 00, 1–8
https://doi.org/10.1093/ccc/tcae029
Original Article