Inequalities and Challenges in Representing Ethnicity Data in Wikidata

MICHELLE NWACHUKWU, King's College London, United Kingdom
MARTIM BRANDÃO, King's College London, United Kingdom
ALBERT MEROÑO-PEÑUELA, King's College London, United Kingdom

Wikidata is the largest open-sourced Knowledge Graph, continuously curated by a global community, and it plays a crucial role in providing structured information for various applications. Its community-driven model allows for individuals from a wide range of diverse backgrounds and knowledge systems to contribute to the growing global knowledge base. However, recent studies have shown that most contributors of Wikidata come from the Global North. Despite a growing awareness of these issues, research on how Wikidata's internal processes shape the structure and representation of ethnicity data remains limited. Where previous studies have focused on queer identity and historical use cases, this paper addresses the gap by systematically analysing the relationship between community discussions and Wikidata content. Using a mixed-methods approach to analyse talk pages (web pages where editors discuss content) and Wikidata graph content, we focus on three properties: Ethnic Group (P172), Country of Citizenship (P27), and Tribe (P12011). This research investigates potential inequalities in the quantity and quality of ethnic-identity data in Wikidata, as well as the challenges editors encounter when working with such data. We find geographical biases in both talk pages and property usage, though these biases vary by property. The challenges editors face when using these properties stem from historical and contextual nuances, legal vs social identity, and issues with ambiguous definitions. Our findings highlight the inequalities, challenges, and the ongoing discourse surrounding ethnic identity related data. However, due to the constraints of Wikidata's structured database, there are built-in biases that have yet to be sufficiently addressed.

CCS Concepts: • **Information systems → Wikis**; • **Human-centered computing → Collaborative and social computing**.

Additional Key Words and Phrases: Wikidata, knowledge graphs, ethnicity, identity, bias

ACM Reference Format:
Michelle Nwachukwu, Martim Brandão, and Albert Meroño-Peñuela. 2026. Inequalities and Challenges in Representing Ethnicity Data in Wikidata. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3805689.3812273

1 Introduction

In recent years, knowledge graphs (KGs) have increasingly been utilised as a tool for AI systems, particularly for modelling, storing, and reasoning over knowledge [17, 25]. Representing real-world entities as a structured, interconnected graph offers a semantic framework that supports context-aware AI systems [29]. The potential of KGs has led to wide adoption in diverse applications, such as recommendation systems [36, 55], large language models [19, 33, 41], and robotics [4, 43, 57].

Authors' Contact Information: Michelle Nwachukwu, King's College London, London, United Kingdom, michelle.nwachukwu@kcl.ac.uk; Martim Brandão, King's College London, London, United Kingdom, martim.brandao@kcl.ac.uk; Albert Meroño-Peñuela, King's College London, London, United Kingdom, albert.merono@kcl.ac.uk.

This work is licensed under a Creative Commons Attribution 4.0 International License.
FAccT '26, Montreal, QC, Canada
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812273

Wikidata is the largest open-source knowledge graph, providing structured data for Wikimedia projects such as Wikipedia. Its community-driven nature can enable diverse perspectives in curating information across interconnected topics. Wikidata significantly shapes online knowledge, raising questions about how knowledge is represented, whether its structure affords diverse views, and whether knowledge from diverse groups is included. The goal of this paper is to characterise such knowledge affordances and inequities on Wikidata, focusing on a type of data tied to diversity, politics, and discrimination: ethnic identity-related data.

Wikidata's ethnicity and country of citizenship data is increasingly used for research, such as racial equity analysis in health data [35], compiling names in African languages [49], or visualising migration of culturally significant people [20]. However, such analyses could be compromised by biases and barriers in contributing ethnicity-related data [7]. Surveys show Wikimedia contributors are predominantly male and from the Global North [61], with downstream effects on content [32]. Despite growing recognition of these concerns [9, 14], limited research examines challenges editors face in documenting identity data, or whether demographic inequalities also translate to content inequalities.

Our goal is to address this gap. More concretely, we ask: **RQ1:** Are there inequalities in the distribution of ethnic identity-related data in Wikidata? We call this the content "quantity gap". **RQ2:** Are there inequalities in the quality of this data, specifically in terms of amount of errors made? We call this the content "quality gap". And finally **RQ3:** What challenges do Wikidata editors face when adding ethnic identity-related data? We go beyond prior work by analysing Wikidata both in terms of graph content and editor discussions. We focus on ethnicity-related properties used to classify Wikidata entities—Country of Citizenship (P27), Ethnic Group (P172), and Tribe (P12011)—using quantitative and qualitative methods. Our results show large inequalities in both the quantity and quality of all three types of data, and highlight contribution challenges tied to historical and contextual nuances, legal vs social identity, and ambiguous definitions. Our work extends research on the difficulties of encoding diverse identities in Knowledge Graphs [7, 60], and on how editor communities shape data quality and ontology design [34, 54, 63]. Such understanding is crucial for developing inclusive AI systems leveraging Knowledge Graphs. We also discuss avenues for future research to overcome these challenges.

2 Background

2.1 Knowledge Graphs

Knowledge Graphs (KGs) are structured representations of data used to convey real world knowledge, consisting of nodes that represent real world entities, and edges that represent relationships between nodes [25]. For example, a KG for books may contain nodes to represent *"titles"*, *"authors"*, *"illustrators"*, *"characters"*, and *"publishers"*; and have edges to represent the relationship between these nodes, such as *"authored_by"*, *"illustrated_by"*, *"character_in"*, and *"published_in"*.

A key benefit of using KGs is that they can be understood by both humans and machines, enabling interoperability between software agents and ease of interaction between KGs and humans. Their structure is designed to be dynamic, continuously updated, shared, and reused. Compared to other traditional database systems, KGs have the advantage of enabling access to multiple sources, or linking different sources together. KGs are used in a wide range of applications such as querying-answering, semantic search, Natural Language Processing, and data integration [29], and they are increasingly used by Big Tech companies in their products (e.g. Google [1], Amazon [15], Airbnb [50]).

2.2 Wikidata

Wikidata is a sister project of Wikipedia and is the largest open source Knowledge Graph [59]. Launching in 2012, it was created and is maintained by a community of volunteers [51]. Wikidata contains over 100 million items and has undergone over 2 billion edits since its launch [3]. In Wikidata, nodes are called *entities*, and edges

are called *properties*. Both properties and entities are assigned unique, machine-readable identifiers (Appendix A). This ensures that a specific concept can be precisely referenced, resolving ambiguity. Q-identifiers are assigned to entities (e.g. "Q5" is "human"), while P-identifiers are assigned to properties, (e.g. "P172" is "Ethnic Group").

Volunteers who contribute to Wikidata are called editors, and there are currently over 25,000 active editors [3]. Talk Pages are where editors collaborate by discussing issues, concerns, and updates. These pages are available on all entity and property pages, allowing any editor to start a discussion. Property Proposal pages are created when an editor suggests a new property. The proposal begins with a rationale for the new property, followed by a discussion where other editors share their support or objections. Ultimately, a decision is made on whether or not to create the new property.

2.3 Contextual Distinction

For the purpose of this paper, we make a distinction between the terms "Western" and "Global North". We use the definition of "Western" from Hall [24], which serves as a historical and cultural descriptor encompassing a lineage of shared philosophies and a history of European colonial expansion. We use the definition of "Global North" and "Global South" from Dados and Connell [13], as an economic and geopolitical categorisation, encompassing wealth accumulation and dominance within global capitalism. This distinction is important due to countries, such as Japan and South Korea, that are categorised within the Global North due to their economic status, but fall outside the cultural and historical classification of the West.

3 Related Work

3.1 Inequity and Power in Wikidata

In 2012 Graham [21] highlighted the issues that could occur from a standardised knowledge base, raising concerns about the standardisation of knowledge and reduction in diversity of perspectives. The article highlighted concerns that this approach could further empower dominant voices in knowledge production, potentially marginalising less-represented groups. Our paper follows the trajectory of these early concerns, analysing data and on-platform editor discussions regarding ethnicity-related data on Wikidata.

Wikidata is designed to be multilingual and officially states its intent to be "an international and thus multilingual project" that is "useful for, users of every language"[1]. Consequently, it is widely treated as a global resource within the research community [12, 30, 31, 42, 49]. Despite these foundational goals, various forms of inequality remain known issues across Wikimedia projects. In a 2021 study, the Wikimedia Foundation surveyed 872 Wikidata editors to understand community demographics and guide future metrics [61]. According to Wikidata statistics[2], at that time there were around 12,000 active editors, of which 75% were male, 16% female, and 2.9% non-binary, 76% from the Global North and 16% from the Global South [2]. The small group of experts that contribute to Wikidata scientific research are also largely from the Global North, as are most funders of the project [58]. Examining race and country of citizenship bias, Shaik et al. [53] found an over-representation of Europe and North America for the professions scientists, software developers, and engineers. Despite increased connectivity, internet access alone does not equate to equitable participation in knowledge production [22], as much of the world remains underrepresented due to a mix of socio-political, economic, and infrastructural barriers.

Wikidata's semantic richness creates a powerful tool for knowledge integration, but its categorisation system can enforce narrow definitions [28], acting as a gatekeeper of information and shaping how knowledge is represented, obscured and shared globally. Baker and Mahal [7] further explores how Wikidata's centralised approach to knowledge and standardised classification reflect particular ideologies. These ontological decisions

[1] https://www.wikidata.org/wiki/Help:Multilingual
[2] https://stats.wikimedia.org/

impact not only what gets represented, but also how knowledge is discussed, where cultural and political assumptions are masked as technical decisions.

This conflation of neutrality and objectivity has led to calls to decolonise structured data. Initiatives such as "Whose Knowledge?" [38] question dominant epistemologies by asking: whose knowledge is structured, and who has the power to curate it? The "Decolonizing the Internet's Structured Data" conference[3] addressed this challenge, highlighting the friction between usability and social justice[46]. Our research is a step in bridging this gap, providing deeper understanding of inequalities in Wikidata ethnicity data and the challenges editors face when contributing.

Research into challenges of documenting identity data is limited. Weathington and Brubaker [60] examines how queer identities are represented in Wikidata, noting cultural biases, stigmatisation risks, and the difficulty of encoding diverse queer identities within structured systems. Baker and Mahal [7] explores the complexity of documenting historical figures into Wikidata and the limitations of strict classification. A Wikipedia study [44] examined biases through a Black feminist perspective, noting how a predominantly White, male editor base and strict reliability standards contribute to racial and gender inequalities. It discusses an assignment involving Black female students improving Wikipedia content, highlighting the potential of diverse contributors while noting the challenges posed by systemic gatekeeping and existing norms. Another Wikipedia study [37], analysed debates in the Italian, Catalan, and French editions, showing how policies and community norms reproduce gender bias against women, non-binary and trans people, limit visibility, reinforce exclusionary knowledge structures, and reflect power hierarchies that favour established editors.

Wikidata operates as a social and collaborative system, where contributors maintain the platform according to community-defined rules, coordinating via talk pages. Various studies have inquired about the impact of discussions within the Wikidata community, both as a way to design more collaborative practices and tools [34], and as a way to assess data quality [54]. One tool for data quality are constraints, rules to ensure consistency. While generally effective, many constraints are outdated or inconsistently applied [54], highlighting the challenge of maintaining structured knowledge at scale.

3.2 Ethnicity data

There is a dichotomy that defines civic and ethnic citizenship as different notions [47]. Whilst civic citizenship stresses law-abidance, ethnic citizenship focuses on ancestry. These concepts vary across different regions and societies. Reijerse et al. [48] extends this research by examining how ethnic, cultural, and civic citizenship interplay and differ through societal perceptions.

Mickel [40] suggest considerations for categorising race and ethnicity in AI fairness work. They note practices are mainly US centric, lacking contextual nuances, and call for critical examination of representation. They highlight issues, such as context-specific categorisation and the risk of misrepresenting and erasing cultures. Jaime and Kern [27] investigates how typical racial categories used in AI fairness literature are not always appropriate, and how ethnicity is a multi-faceted concept with overlapping measures forming identity. These are also aspects that we explore in this paper, through analysis of inconsistencies and editor challenges in the use of ethnicity-related data on Wikidata.

4 Methodology

This study adopts an interdisciplinary methodology that integrates computational data analysis of Wikidata entries and thematic analysis of Wikidata Talk pages. The mixed-methods approach is used to analyse potential biases in the formulation, discussion, and use of ethnicity-related identifiers in Wikidata.

[3] https://whoseknowledge.org/wp-content/uploads/2022/04/DTI-SD-SummaryReport-ENGLISH.pdf

4.1 KG Data Collection and Analysis

We used QLever [8], a querying service that supports SPARQL (the RDF query language), for data extraction. A SPARQL query was developed to retrieve relevant properties associated with ethnic identity. We first identified Ethnic Group (P172) as the property most related to ethnicity, and then ran a SPARQL query using the property "Related Property" (P1659) and identified Country of Citizenship (P27) and Tribe (P12011) as relevant properties to be analysed.

Further SPARQL queries were developed to count the number of people (Q5) associated with the identified properties. We further grouped them into associated countries (P17) and continents (P30), where applicable, for comparative analysis. We did this by first categorising people based on their Country of Citizenship (P27), Ethnic Group (P172), and Tribe (P12011). We then ran a second query to link each entity to its corresponding country (P17) and continent (P30) in Wikidata (e.g., ethnic group -> P17 -> country), effectively assigning geographic context to each identity. We also gathered property constraints (P2302) to assess the quality of the data, in terms of number of entries which do not satisfy property constraints.

For the purpose of this research, we counted instances of property usage rather than distinct individuals. While this approach results in an aggregated count that exceeds the number of unique human entities (due to 8.3% of individuals possessing multiple citizenships or holding citizenships that varied over time) this was a deliberate choice. We avoided selecting a primary or recent citizenship for a person as this would be a form of data reduction that flattens intersectional identities and has the potential to introduce researcher bias.

As part of our analysis, we compared the instances in the KG with national population and internet population data. We gathered national population data from the CIA World Factbook database[4], and internet population data from the International Telecommunication Union[5]. There is a slight discrepancy in this analysis as we compare distinct individuals with Wikidata instances. As a result, the representation ratio for certain nations in the KG may be marginally elevated by individuals holding multiple citizenships. However we feel that this approach accurately captures the cultural and historical representation of a region and reflects its footprint within Wikidata.

4.2 Talk Page Data Collection & Analysis

We also gathered text data from the associated property Talk Pages for Country of Citizenship, Ethnic Group, and Tribe. There is currently no property for nationality, so in addition, we analysed discussions for property proposals related to notions of nationality: *"Nationality"*, *"Cultural Identity"*, and *"Nationality (cultural identity)"*. We included these in our study, along with property proposal pages of Country of Citizenship and Tribe, to ensure that the broader context of how these terms are conceptualised and defined are considered and how each property and suggested property interplay with each other. The property proposal page for Ethnic Group could not be located. A list of the URLs for the 8 primary pages we analysed can be found in Appendix F. We also selected a set of archived discussions by conducting a keyword search for: "citizenship, ethnic group, ethnicity, tribe, nationality, cultural identity." From these initial results, we manually reviewed and selected relevant pages that directly discussed these concepts.

In order to identify trends of ethnicity-related properties, we also conducted frequency analysis and counted the number of times territories, or ethnicities related to territories, occurred in talk pages. Each talk page was split into sections, where a section was a new topic of discussion identified by a heading written by the editor that brought forward this topic of discussion. This allowed us to see which countries are most frequently referred to when giving examples during discussions. Each page was scraped and territories were counted using a Python script.

[4]https://www.cia.gov/the-world-factbook/field/population/
[5]https://datahub.itu.int/data/?i=11624

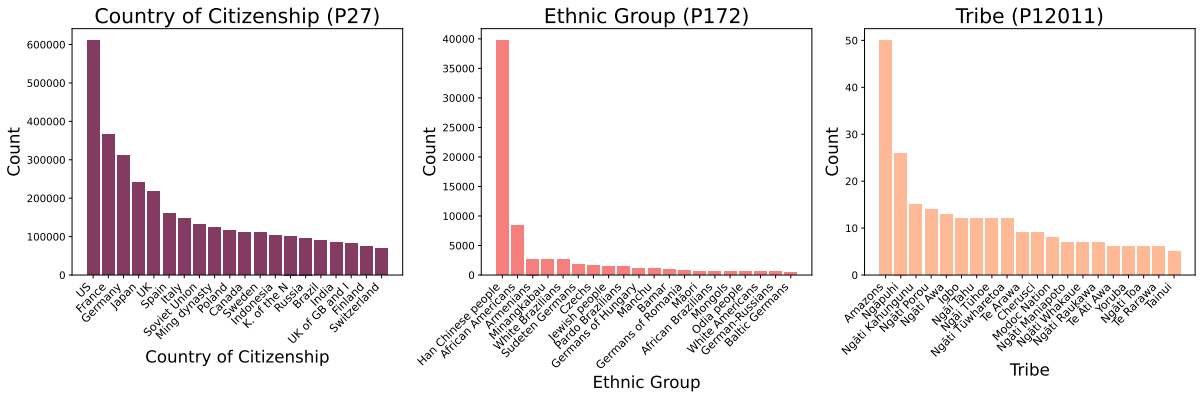

Fig. 1. Top 20 SPARQL query results for the properties "Country of Citizenship", "Ethnic Group", and "Tribe", illustrating usage distribution and highlighting disparities in representation. *"US" - United States, "UK" - United Kingdom, "K. of the N" - Kingdom of the Netherlands, "UK of GB and of I' - United Kingdom of Great Britain and of Ireland*

We used an inductive thematic analysis approach [11]. To ensure reliability, two researchers collaboratively went through the analysis process by first independently coding the talk pages. We met regularly to iteratively refine the codebook and discuss coding decisions. In this way we were able to reach agreement on the themes, focusing on identifying recurring patterns of discussions related to challenges in the use of properties.

5 Results: Content Quantity Gap

The analysis of SPARQL query results highlights an imbalance in the usage of the Wikidata properties (Figure 1). Of the three properties analysed, Country of Citizenship is the most used, with over 5 million total entries (assignments of people to countries) and more than 3,200 distinct entities (distinct countries). Ethnic Group has far fewer total entries, approximately 84,000, but still includes around 2,200 distinct entities. The United States accounts for the most Country of Citizenship entries. Among the top seven countries, Japan is the only Asian country represented, while the others are the United States and Western European countries.

Han Chinese people is by far the most represented Ethnic Group (Figure 1). African Americans greatly outnumber White Americans. Broad racial categories such as Black people and White people are also present, raising questions about the appropriateness of these labels (Appendix B). Discussions on Wikidata have touched on the controversial nature of adding race as an ethnicity, which may contribute to these inconsistencies. Overlapping labels such as British and English suggest differing interpretations or labelling practices for the same group, potentially affecting accuracy.

The Tribe property is far less used, with 582 total entries and 250 distinct entities. Its usage presents certain challenges. Some groups, such as Yoruba and Igbo, are used inconsistently as both ethnic groups and tribes, highlighting overlaps in categorisation that may affect reliability. Certain tribes appear under multiple names (e.g., Yoruba and Yoruba people), creating ambiguity in entity usage. For instance, Yoruba (Q34311) refers to the language entity, while Yoruba people (Q190168) refers to the ethnic group entity.

It is worth noting that Wikidata has the capacity to express multiple ethnicities and citizenships through its taxonomy, allowing for the representation of various aspects of a person's identity by allowing for multiple Ethnic Group and Country of Citizenship entries. However, it is rare for both Ethnic Group and Country of Citizenship to be used together for a single individual.

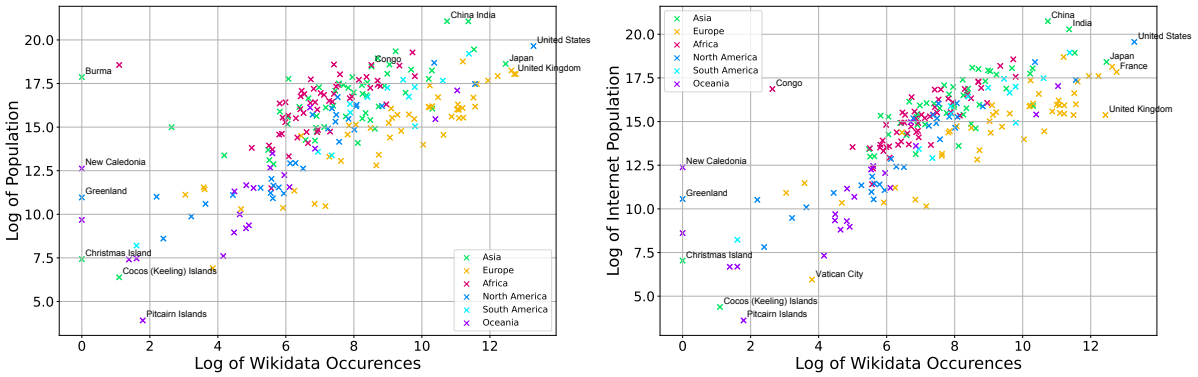

Fig. 2. Relationship between Wikidata "Country of Citizenship" occurrences and population. Log-log plots compare each country's number of occurrences with its national (left) and internet population (right). Colours indicate continental regions. Both plots reveal positive linear trends, suggesting power-law relationships.

When results are examined by continent using SPARQL queries (Figure 6), Europe has by far the most entries for Country of Citizenship, followed by Asia. For Ethnic Group, Asia leads (dominated by Han Chinese), followed by North America and Europe. The Tribe property, by contrast, has far fewer entries and shows a different pattern with Insular Oceania, Oceania, the Australian continent, and Africa leading, while Europe has comparatively few. Not all Countries of Citizenship, Ethnic Groups, and Tribes, have an associated country or continent documented on Wikidata, so these graphs represent only a subset of Wikidata.

5.1 Comparison between Population and Wikidata Contents

5.1.1 Country of Citizenship. Figure 2 shows the relationship between Country of Citizenship occurrences and both national population and internet populations across countries. We used log-log plots to better capture and interpret this relationship. The resulting plots show positive linear trends, indicating that countries with larger populations generally have more Wikidata occurrences.

All graphs, overall and by continent, show statistically significant positive linear relationships (Pearson correlation, $p<0.001$). Overall trends are sublinear (gradients<1) or near-linear (gradients≈ 1), with slight slope variations between continents. Notably, Africa's internet population plot shows a stronger sublinear relationship, suggesting slower growth in Wikidata occurrences relative to internet population increase.

For all continents combined, the coefficient of determination (R^2) indicates moderate fit: 0.64 for national population and 0.50 for internet population. By continent, the fit improves, with strong to very strong R^2 values (0.75-0.89). Exceptions include Oceania (moderate fits), Asia (moderate fit for internet population), and Africa (weak fits, very weak R^2 of 0.22 for internet population). European countries consistently have more Wikidata entries than other continents, suggesting under-representation elsewhere. Despite very large populations, China and India have fewer entries than smaller populations, such as the United States, the United Kingdom, Germany, France and Japan. Some countries with seemingly large populations, like Burma, also have low occurrences.

These results suggest that population, particularly national population, is a good predictor of Wikidata occurrences. However, the relationship varies by continent with many outliers, and Europe being an exception where internet population is a better predictor than national population.

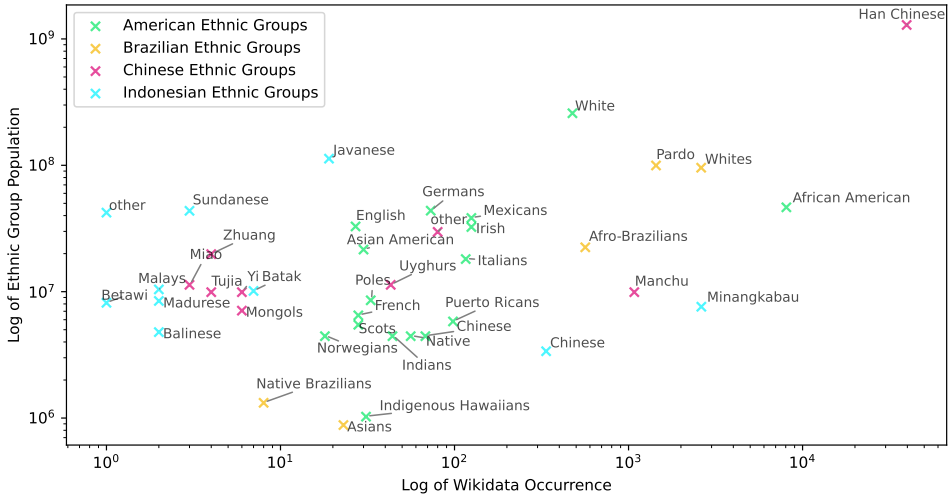

Fig. 3. Relationship between Wikidata "Ethnic Group" property occurrences and population. The log-log plots compares the number of times each ethnicity appears in Wikidata with its corresponding population within a country. Colours represent different countries.

5.1.2 Ethnic Group. Next, we investigated the relationship between ethnic group occurrences on Wikidata and the groups' national population. We selected China, America, Brazil and Indonesia due to their ethnic groups having the highest occurrence on Wikidata, and availability of reliable data[6].

We now summarise the patterns we identified, though a plot is also shown in Figure 3. Different patterns emerge related to the dominant ethnic groups. In China, "Han Chinese" has far more occurrences than others. However, this matches its population statistics, since Han Chinese accounts for 91% of China. In America and Indonesia, the pattern is different, since dominant ethnic groups are under-represented. For American ethnic groups, "White" has a very low Wikidata occurrence compared to its population, and "African American" is more represented despite a smaller population. In Indonesia, "Minangkabau" has high Wikidata occurrence despite a low population, and the dominant Javanese ethnic group has relatively few entries. Brazil presents a third, different pattern. "Pardo" is an ethnic term and skin colour classification but more commonly used to describe those of mixed heritage. There is roughly an equal population of Pardo and White people in Brazil, but on Wikidata, "White Brazilians" have higher occurrence, unlike the American pattern where White is under-represented.

In the discussion, we will explore how these patterns reflect different practices in the use of the "ethnic group" property and varying conceptions of ethnicity.

5.2 Examples in Talk Pages

We next analysed the number of times a country or territory was mentioned in Talk Page discussions. This included 169 sections where there were 358 mentions of countries, territories and ethnicities. 90 countries were recorded from this analysis. The top 8 countries mentioned in the property talk pages, led by the United States, Germany, and the United Kingdom, are part of Western Europe and the United States and account for a significant proportion of mentions (Figure 4). They total 173 mentions which is 48% of the total countries mentioned. This suggests a bias in the focus of Wikidata discussions. Figure 5 shows the data split according to their geographical

[6]https://www.cia.gov/the-world-factbook/field/ethnic-groups/

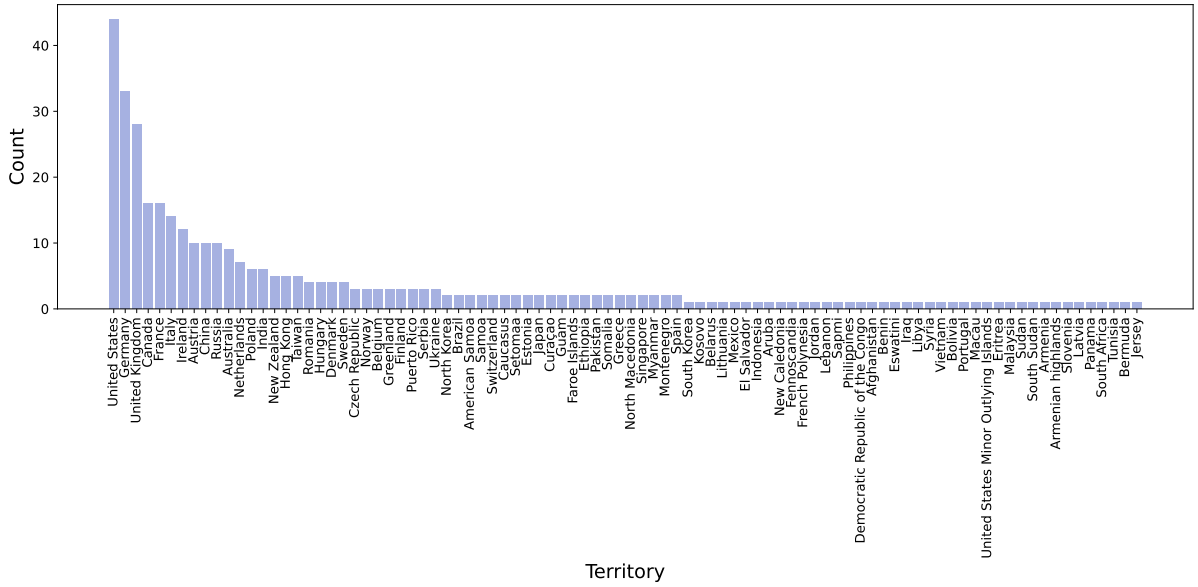

Fig. 4. Territories mentioned in Wikidata property discussions. The United States, Germany, and the United Kingdom lead in frequency, with the top eight countries accounting for 48% of all country mentions, indicating a strong geographical skew.

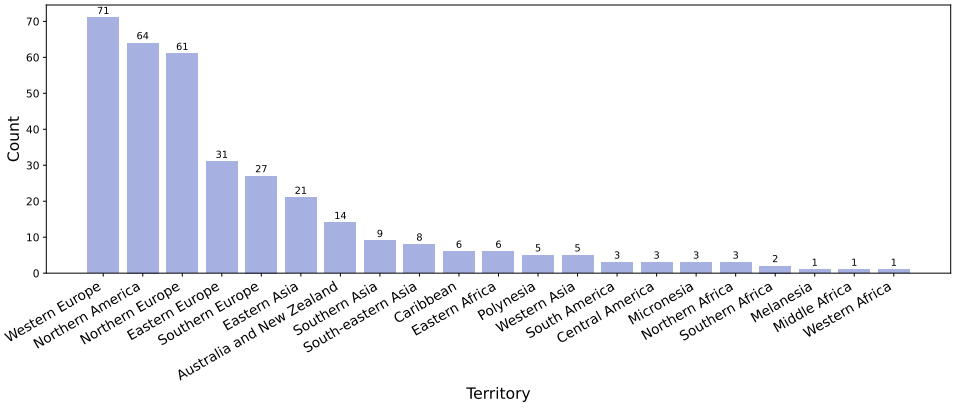

Fig. 5. Mentions of territories in Wikidata property discussions, grouped by geographical subregion. The figure shows a strong skew toward Northern America and Europe, which together account for 71% of total mentions.

subregion. The figure shows a skew towards Northern America and Europe, of which they account for 71% of the total mentions.

6 Results: Content Quality Gap

As part of how each property is defined, property pages contain a list of constraints which state the type of entity the subject and object can be, in any triple that uses the property (where subject -> property -> object,

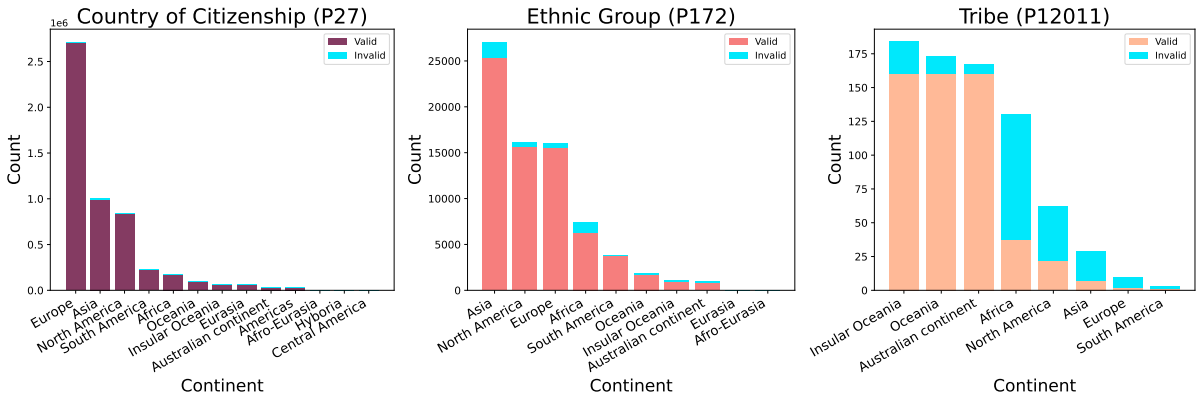

Fig. 6. SPARQL query results showing property usage by continent. Stacked bar plots display the number of valid and invalid triples for each property. Invalid triples are those that violate defined constraints. The figure reveals distinct patterns: "Country of Citizenship" is predominantly used by European countries, "Ethnic Group" by Asian countries, and a large proportion of "Tribe" triples are invalid.

is a triple). A constraint not met is technically a "mistake", and thus an indication of low data quality. As part of our investigation we looked at all triples where human (Q5) is the subject type and found where the object constraints are not met. For example Country of Citizenship has constraints such as *'political territorial entity'*, *'nationality'*, *'sovereign state'*, *'citizenship'*, *'country'*; Ethnic Group has constraints such as *'indigenous people'*, *'tribe'*, *'nationality'*, *'ethnic minority group'*, *'human race'*; and Tribe has constraints such as *'Berber tribe', 'Native American tribe'*, *'tribe'*, *'iwi'*, *'Seven Slavic tribes'*.

Figure 6 shows both valid and invalid triples per continent, where an invalid triple is one that does not lie within the constraints (e.g. assigning Yoruba language to an ethnic group). Despite Europe having by far the most Country of Citizenship entries, Asia has the most invalid triples according to the constraints. Among all the continents documented on Wikidata, Asia has the highest percentage of invalid triples at 1.33%, followed by Africa with 1.17%. The percentage of invalid Country of Citizenship triples is comparatively lower when compared to Ethnic Group and Tribe. For Ethnic Group, Europe, North America, and South America have a lower percentage of invalid triples compared to other continents. Tribe, on the other hand, has a large percentage of invalid triples, which may indicate a lack of usage and refinement of this property compared to the others.

7 Results: Editor Challenges

After completing the quantitive analysis of Wikidata, we conducted thematic analysis of the talk pages Country of Citizenship (P27), Ethnic Group (P172), Tribe (P12011) and relevant property proposal and archived pages. We identified the following editor challenges.

7.1 Ambiguity and Misuse

There is a common aspect of the discussions that highlights the ambiguity of these properties that can lead to misuse and undermines their accuracy and reliability. Some view Country of Citizenship as being too generic, stating that the complexities of how states evolve over time make this property difficult to use properly: *"[T]his property is just too difficult for many states IMHO (due to wars, merger of countries, sep[a]rations, etc.), it would be some kind-of nightmare to fix this for all persons."* Ethnic Group is often seen as vague and hard to apply, assigning

ethnic groups based on traits or ancestral origins is deemed problematic, and this vagueness causes the systematic misuse of this property.

There are also concerns on the often blurred boundaries between the terms citizenship, ethnicity, and nationality that are often used interchangeably, but potentially that a new Nationality property may solve: *"[T]he lack of understanding [...] on whether this sense of nationality should be recorded in country of citizenship (P27) or ethnic group (P172) seems to indicate that we have a problem that this new property might solve"*

7.2 Legal vs. Social Identity (neutrality vs opinion)

There is a debate within these property proposal pages on whether citizenship should be strictly a legal definition or should allow for other state identifications. Wikidata uses the legal definition for "neutrality"[7][8], but editors note that this leads to exclusion and complexities related to postcolonial contexts, diasporas, and stateless communities: *"The child has Mexican citizenship, Mexican legal nationality, and was raised to believe that s/he had been born in America and encouraged to self-identify as belonging to the US [...] I don't think that the current set of properties allow us to correctly record this situation."*

However, some editors believe that detaching the legality from Country of Citizenship would dilute its integrity: *"I don't think a solution is to remove the legal question from the definition, without it would become a meaningless «catch-all» category"* Many view the lack of legal affiliation as causing the property to lack usefulness or reliability, further exacerbated by ambiguity in laws in countries such as France and Germany, where building such datasets is illegal [5, 26]: *"This property will be as a mess as ethnic group (P172). Also this is [forbidden] at least in France and Germany to label this thing. [...] By law, the countries can't take this information by censuses".*

7.3 Historical and Contextual Nuances

Contributors consistently put forward concerns that the properties Country of Citizenship and Ethnic Group are rooted in modern legal frameworks/national constructs and Western-centric classifications, and are not adaptable for the diverse and context dependent identities of people. Applying these properties for some subjects is often done inconsistently and inappropriately. Country of Citizenship is frequently challenged on its usage for historical figures where either the legal notion of citizenship, or the country itself, did not exist: *"I find this property a bit too generic and useless, [...] the description should really emphasize that this should only be used for persons who really have a citizenship of this modern state (and not any previous state which has been "merged" into that state)."*

The creation of Tribe (P12011) can see the importance of contextual nuance. There is some contention on whether this property is necessary as there is overlap with current properties: *"Oppose [d]epending on the requirements for official recognition or documentation this is redundant with either ethnic group (P172), member of (P463), or country of citizenship (P27)."* But it is established that in some parts of the world, a tribe is not considered to be an ethnic group, so the property was ultimately created: *"Our tribes are not ethnic groups. My Aunt's country of citizenship is New Zealand, her ethnic group is Māori, her Tribe (Or Iwi in the local language), is Ngāti Kahungunu. All different things."*

Similarly, the term "nationality" has a meaning that shifts depending on the culture and historical context: *""[N]ationality" [...] tends to have quite different connotations in Britain and the United States. [...] For example, it is not unusual for an American to say his nationality is Irish if his parents come from Ireland, even if he was born in the U.S."* Nationality can mean cultural belonging, but it can also mean legal citizenship [56]. Because of this, many editors feel apprehensive about creating a new nationality property: *"I think this is a rather slippery concept, likely to be historically and culturally contingent and readily challenged."*

[7] https://www.wikidata.org/wiki/Wikidata:Living_people
[8] https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

7.4 Controversies and Bias Concerns

Documenting personal data is bound to come with impassioned disagreements as lived experiences are discussed and debated. The talk pages provide an account of these disagreements that touch upon issues of neutrality and differences in perspectives that stem from the editor's personal experiences. The lack of clarity and consensus on the use and meaning of these properties has led to conversations around bias, sensitivities, and controversies. As an example, one editor in a discussion on Country of Citizenship criticises the "moderately educated Americans" perspective that influences Wikidata and suggests international sources instead of legal frameworks from a small subset of countries. The user is criticised by another user who says that their interpretation of international law may be linguistically or culturally biased, and warns against imposing one national perspective on an international platform.

Other conversations touch on the notion of a Western framework being the main influence of Wikidata: *"I think both properties should be merged and don't have "modern western contemporary legal nation state membership" in one side and "all the other forms of human organization" in the other."*. There is a particular conversation where an editor points out the misrepresentations of assigning US citizenship to Puerto Ricans born before 1899, and challenges the erasure of colonial and indigenous identities by the strict modern legal citizenship definitions. Another editor defends the current Wikidata model because of the need for data consistency. They acknowledge the flaws in the conceptualisation of modern nation-states but says they provide a usable legal framework for data purposes.

7.5 Structure, Queryability, and Usability

Some discussions touch on the need for clear, queryable properties whilst also attempting to respect the complex nature of human identity. There is some concern about the limitations of SPARQL querying when it comes to structuring this data. A suggestion of using *"novalue"* when recording someone's statelessness identity was deemed not useful due to the *"novalue"* term being invisible to queries. However, avoiding the use of *"novalue"* can lead to inconsistencies in Wikidata's data modelling. Wikidata is built on the open-world assumption (OWA) [6], where missing data is deemed unknown. In practice, many SPARQL queries assume the closed-world assumption (CWA), treating unknown data as false [62].

Editors note that applying Country of Citizenship across historical border changes, or using Ethnic Group across different national definitions, makes consistent querying difficult. This is especially evident when automated systems try to reuse this in a way that is beyond the scope of the data, unintentionally producing misrepresentations of a person's identity: *"Templates like c:Template:Creator are attempting to analyze country of citizenship (P27), ethnic group (P172) [...] and come up with "Nationality" [...] but often it leads to quite comical results. [...] Adam Mickiewicz [is described by] most sources as Polish or Polish/Lithuanian poet, but nothing in the [Wikidata] captures that: his country of citizenship was Russia, [...] he spoke [a] number of languages, and his ethnic group was Belarusian. [...] We need a sourced way to capture this information."*

7.6 Proposed Solutions

Editors are frequently proposing solutions to help clear up some of the confusion that has become synonymous with this set of properties. Suggestions for Country of Citizenship include assigning multiple values with qualifiers like end dates to better reflect political and border changes over time. Others ask to create more generic items to reduce complexity, for example using an overarching UK item: *"[T]here needs to be a generic UK item to link with the Wikipedia articles, and since they are already linked with United Kingdom (Q145), that's probably the right item."*

There is broad agreement that Ethnic Group needs strong sources when attributing ethnic identity which is further confirmed by the definition of the property (Appendix A). This is because of Wikidata policy of sensitivity toward personal data and ethical concerns. Some suggest a mass cleanup and the removal of unsourced claims as

a solution. Others suggest clarifying the scope of Ethnic Group with constraints or value types to avoid merging it with nationality or cultural identity. And there are suggestions of adding nationality as an alias to Country of Citizenship or Ethnic Group, which a version of the alias has been added to both (Appendix A).

A proposed nationality property has been suggested three times for a potential solution. Some propose creating structured items like "German nationality" rather than linking directly to countries. But there are also concerns on the risks of promoting nationalist biases. There is also concern of duplicating Country of Citizenship's purpose so a suggestion of renaming Country of Citizenship as *"country of citizenship or nationality"*. There are also other name suggestions such as *"national identity"* and *"ascribed cultural identity"*.

8 Discussion

Our analysis highlights the epistemological bias embedded in the definition and usage of the Wikidata properties Country of Citizenship (P27), Ethnic Group (P172), and Tribe (P12011). We show this through analysing inequalities in the frequency and quality of ethnic identity-related data, and by contrasting these with common themes among editors, such as ambiguity, neutrality, context, and usability. The representation of countries, territories, and ethnicities in these properties reflects broader global disparities in digital representation, disproportionality catering to Western European and North American perspectives [61]. This skew makes it difficult to use these properties in knowledge systems rooted in other epistemological traditions, limiting the usability and inclusiveness of Wikidata as a global knowledge base. Wikidata is an epistemic artifact shaped by the social and cultural contexts of its creators and dominant contributors. Importance should be placed on the epistemological foundations of Wikidata, rather than trying to conform knowledge to biased constraints. These findings have implications for understanding Wikidata's reliability as a resource and for addressing its structural inequities.

8.1 RQ1&2: Geographical Biases

Our results show there is a clear overrepresentation of Western Europe and the United States in both discussions and property usage. The top eight countries mentioned, led by the United States, Germany, and the United Kingdom, account for nearly half of all country mentions in talk pages, indicating a disproportionate focus on these regions. A similar pattern emerges when analysing subregions. These results likely stem from a combination of historical, economic, and digital inequalities that have led certain regions to dominate global knowledge creation and curation [52]. Countries with large populations but significantly fewer Wikidata entries, such as Burma and the Democratic Republic of Congo, tend to lie outside Europe and the United States. These findings raise concerns about Wikidata's ability to function as a truly global resource. This view is reflected by Ford and Iliadis [18], who describe the inevitable political and social aspects of Wikidata and the power it holds in distributing knowledge online, while also highlighting the loss of cultural nuances.

The imbalance in the usage of Country of Citizenship and Ethnic Group further highlights structural biases within Wikidata. While Country of Citizenship is the most heavily used property out of the three we analysed (dominated by Western countries and Japan), Ethnic Group is relatively underused, with approximately 100 times fewer entries. This suggests that contributors may prioritise certain types of identity information over others, possibly due to cultural or contextual factors. Such patterns are evident in the contrast between the top countries for Ethnic Group, such as China, Brazil, and Indonesia, and those for Country of Citizenship. The tribe property sees minimal usage, with an average of about two entries per tribe. This reflects both low engagement and challenges in applying the term consistently. Confusion between tribal and linguistic categories, such as the overlap between Yoruba as a tribe and Yoruba as a language, further highlights the need for clearer definitions and standardised usage.

8.2 RQ3: Controversies and Ambiguity

Our results show that discussions in Wikidata reflect a tension between creating a usable KG and creating one that expresses the complexity of human identity—a core challenge of Wikidata [46]. It is common for editors to be put forward examples that influence property definitions in order to better express these complexities. However, these examples overwhelmingly come from Europe and North America, reflecting the dominance of these regions in Wikimedia participation [61]. The frustration towards misrepresentation is evident throughout the discussions and is consistent with prior research investigating identity representation in Wikidata [7, 60].

In the discussions, it was mentioned that Country of Citizenship aligns more closely with Western notions of citizenship, while Ethnic Group captures identities that do not fit within these frameworks. This may explain why Ethnic Group is more commonly used for Asian entities, as it better reflects their identities. Although Asia has the second-highest number of Country of Citizenship entries, this is less surprising given its large population. However, it may still indicate underrepresentation, particularly as North America follows closely in third place despite having a smaller population.

These patterns reflect the dominance of Western-centric definitions of identity, which often exclude other contexts and experiences. Other studies looking at identity data in Wikidata [7, 60] have not explicitly investigated the dominance of Western perspectives. A common thread in similar studies and critiques is the risk posed by neutrality and centrality, foundational pillars of Wikidata, producing reductive representations that conceal and reinforce biases [7, 18, 21, 60]. This study contributes to the discourse through a quantitative analysis of Wikidata entries, offering insights that strictly qualitative studies may overlook.

8.3 Ethnic Group Usage Challenges and Inconsistencies

Han Chinese is by far the most represented ethnic group, likely driven by the inclusion of historical figures and potentially automated contributions [23]. The contrast between the top countries represented by Country of Citizenship and those represented by Ethnic Group reflects differing patterns of usage.

Inconsistencies also appear within countries. For example, African Americans are more frequently represented than White Americans. This may reflect community specific engagement with Wikidata [44], or differing attitudes towards documenting race and ethnicity [16]. Within the United States, where White Americans are the majority demographic, use of this property may be less common as "White" is often perceived as the normative identity, reducing the likelihood of explicit categorisation [16, 45]. This is evident in the frequent absence of ethnicity data for White Americans [39].

These inconsistencies distort the dataset, creating the impression that minority groups such as African Americans are overrepresented, while simultaneously obscuring the broader overrepresentation of White Americans across both general and professional categories [39, 53]. Another contributing factor may be the use of more specific ethnic labels such as "German American" or "Italian American".

In contrast, connotations with categorising someone as white within America may be different to categorising someone as white within Brazil as although "White" is not the majority ethnic group in Brazil, it is more frequently recorded than other ethnic groups in Wikidata. This may reflect the same underlying mechanism: ethnic identities are more likely to be recorded when they deviate from the perceived norm. In Brazil, where "White" is not the default category, it is therefore more often explicitly documented. The use of this property can, in turn, contribute to the othering of certain communities.

The data also reveals inconsistencies in how ethnicities and tribes are categorised. Overlaps exist between categories, for example "Yoruba" is classified as both an ethnic group and a tribe, or labels that blur distinctions such as "British" and "English". Broad racial categories such as "Black people" and "White people" introduce additional challenges as they often fail to capture the specific cultural or social dimensions of an individual's identity. These issues are compounded by the fact that people are racialised according to specific cultural contexts,

meaning that identity labels can vary depending on who is applying them and from which cultural standpoint [10]. Such conceptual ambiguities have practical consequences as they contribute to inconsistent application of identity properties. Previous analyses have reported that nearly half of the statements using Ethnic Group violate property constraints [7]. From our investigation we found this to not be the case (Figure 6) and this may be the result of regular mass deletions which are referenced in the talk pages. These inconsistencies highlight the need for clearer guidelines on how ethnicities and tribes should be represented in Wikidata, in order to reduce ambiguity and enhance data quality.

9 Conclusion

In this paper we have highlighted the inequalities and complex challenges editors face when representing ethnicity data within Wikidata's structured database, specifically investigating the properties Country of Citizenship (P27), Ethnic Group (P172), and Tribe (P12011). While rules and platform-wide etiquettes aspire to inclusivity and global knowledge representations, we find geographical biases in both talk pages and property usage. We also highlight challenges stemming from historical and contextual nuances, legal vs social identity, and ambiguous definitions.

Limitations This work has several key limitations. First, the perspectives of the "silent majority" of editors (the majority of active editors who make no contributions to talk pages) are not captured in the talk page analysis and therefore cannot fully represent the views of all Wikidata contributors. Second, we restricted our scope to English-language content, which does not account for how bias manifests in non-Western or under-resourced language editions; talk page discussions in other languages are limited or, in many cases, non-existent. Third, our research did not account for individuals with multiple citizenships or citizenships that varied over time, meaning we did not analyse these intersectional nuances.

Future Work To move toward more equitable practices in online knowledge curation and exchange, we propose several directions for future research. 1) Restructure and rename ambiguous and overlapping properties to reflect a more nuanced understanding of identity. For example, adding and renaming properties to *"nationality as country of birth"*, *"nationality as cultural identity"*, and *"nationality as citizenship"* as subproperties of a broader nationality class could allow for greater coherence and inclusion to cover different notions of identity. This model would allow queries to return either specific subproperties or the overarching property. 2) Develop tools to support editors in identifying and flagging incorrect or outdated uses. These flags could then be reviewed to determine whether the instance should be corrected, or whether the property itself requires updating i.e. adding new constraints. 3) Design and deploy systems to help editors understand the real-world consequences of categorical classifications. Future work should include creating visualisation tools that make Wikidata content transparent, as well as bias-detection frameworks that assess how AI systems using Wikidata might perpetuate or amplify inequities.

Acknowledgments

This work was supported by the UK Research and Innovation Centre for Doctoral Training in Safe and Trusted Artificial Intelligence [9] [grant number EP/S023356/1].

Generative AI Usage Statement

ChatGPT was used for grammar corrections and improving the fluency of writing. All content was reviewed and verified by the authors, and revised to ensure accuracy and authenticity of the intended meaning.

[9]https://www.safeandtrustedai.org

References

[1] 2012. Introducing the Knowledge Graph: things, not strings — blog.google. https://blog.google/products/search/introducing-knowledge-graph-things-not/. [Accessed 13-05-2024].

[2] 2024. Economy classifications — itu.int. https://www.itu.int/en/ITU-D/Statistics/Pages/definitions/regions.aspx. [Accessed 13-05-2024].

[3] 2025. Wikidata:Statistics. https://www.wikidata.org/wiki/Wikidata:Statistics. [Accessed 22-05-2025].

[4] Mark Adamik, Ilaria Tiddi, Romana Pernisch, and Stefan Schlobach. 2024. Large-Scale Knowledge Graphs as a Tool for Enhanced Robotic Perception. In *Proceedings of the AAAI Symposium Series*, Vol. 4. 338–345.

[5] Pierre Affagard and Laure Ekani. 2025. Data Protection Laws and Regulations Report 2025 the increased relevance for companies of data collection of racial and ethnic origins in the EU. https://iclg.com/practice-areas/data-protection-laws-and-regulations/03-the-increased-relevance-for-companies-of-data-collection-of-racial-and-ethnic-origins-in-the-eu?utm_source=chatgpt.com

[6] Hiba Arnaout, Simon Razniewski, Gerhard Weikum, and Jeff Z Pan. 2021. Negative knowledge for open-world Wikidata. In *Companion Proceedings of the Web Conference 2021*. 544–551.

[7] James Baker and Ammandeep K Mahal. 2024. "I have always found the whole area a minefield": Wikidata, historical lives, and knowledge infrastructure. *International Journal of Digital Humanities* (2024), 1–20.

[8] Hannah Bast and Björn Buchhold. 2017. Qlever: A query engine for efficient sparql+ text search. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 647–656.

[9] Bettina Berendt, Oğuz Özgür Karadeniz, Sercan Kıyak, Stefan Mertens, and Leen d'Haenens. 2023. Diversity and bias in DBpedia and Wikidata as a challenge for text-analysis tools. *o-bib. Das offene Bibliotheksjournal/Herausgeber VDB* 10, 2 (2023), 1–12.

[10] Geoffrey C Bowker and Susan Leigh Star. 2000. *Sorting things out: Classification and its consequences*. MIT press.

[11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[12] Ruixiang Cui, Rahul Aralikatte, Heather Lent, and Daniel Hershcovich. 2022. Compositional generalization in multilingual semantic parsing over Wikidata. *Transactions of the Association for Computational Linguistics* 10 (2022), 937–955.

[13] Nour Dados and Raewyn Connell. 2012. The global south. *Contexts* 11, 1 (2012), 12–13.

[14] Paramita Das, Sai Keerthana Karnam, Anirban Panda, Bhanu Prakash Reddy Guda, Soumya Sarkar, and Animesh Mukherjee. 2023. Diversity matters: Robustness of bias measurements in Wikidata. In *Proceedings of the 15th ACM Web Science Conference 2023*. 208–218.

[15] Xin Luna Dong. 2018. Challenges and innovations in building a product knowledge graph. In *Proceedings of the 24th ACM SIGKDD International conference on knowledge discovery & data mining*. 2869–2869.

[16] Richard Dyer et al. 2005. The matter of whiteness. *White privilege: Essential readings on the other side of racism* 3 (2005).

[17] Lisa Ehrlinger and Wolfram Wöß. 2016. Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)* 48, 1-4 (2016), 2.

[18] Heather Ford and Andrew Iliadis. 2023. Wikidata as semantic infrastructure: Knowledge representation, data labor, and truth in a more-than-technical project. *Social Media+ Society* 9, 3 (2023), 20563051231195552.

[19] Yanjun Gao, Ruizhe Li, Emma Croxford, John Caskey, Brian W Patterson, Matthew Churpek, Timothy Miller, Dmitriy Dligach, and Majid Afshar. 2025. Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *JMIR AI* 4 (2025), e58670.

[20] Arjun Ghosh. 2024. Visualizing the Cultural History of South Asia. In *Practices of Digital Humanities in India*. Routledge India, 69–95.

[21] Mark Graham. 2012. The Problem With Wikidata. https://www.theatlantic.com/technology/archive/2012/04/the-problem-with-wikidata/255564/. [Accessed 18-11-2024].

[22] Mark Graham, Bernie Hogan, Ralph K Straumann, and Ahmed Medhat. 2014. Uneven geographies of user-generated information: Patterns of increasing informational poverty. *Annals of the Association of American Geographers* 104, 4 (2014), 746–764.

[23] Andrew Hall, Loren Terveen, and Aaron Halfaker. 2018. Bot detection in wikidata using behavioral and other informal cues. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–18.

[24] Stuart Hall. 2007. The West and the rest: Discourse and power. *Race and racialization: Essential readings* 56 (2007).

[25] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. 2021. Knowledge graphs. *ACM Computing Surveys (Csur)* 54, 4 (2021), 1–37.

[26] Lars Hornuf, Sonja Mangold, and Yayun Yang. 2023. Data protection law in Germany, the United States, and China. In *Data Privacy and Crowdsourcing: A Comparison of Selected Problems in China, Germany and the United States*. Springer, 19–79.

[27] Sofia Jaime and Christoph Kern. 2024. Ethnic classifications in algorithmic fairness: Concepts, measures and implications in practice. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 237–253.

[28] Krzysztof Janowicz, Bo Yan, Blake Regalia, Rui Zhu, and Gengchen Mai. 2018. Debiasing Knowledge Graphs: Why Female Presidents are not like Female Popes.. In *ISWC (P&D/Industry/BlueSky)*. 1–5.

[29] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems* 33, 2 (2021), 494–514.

[30] Lucie-Aimée Kaffee, Alessandro Piscopo, Pavlos Vougiouklis, Elena Simperl, Leslie Carr, and Lydia Pintscher. 2017. A glimpse into Babel: an analysis of multilinguality in Wikidata. In *Proceedings of the 13th International Symposium on Open Collaboration*. 1–5.

[31] Lucie-Aimée Kaffee and Elena Simperl. 2018. Analysis of editors' languages in wikidata. In *Proceedings of the 14th International Symposium on Open Collaboration*. 1–5.

[32] Ozgur Karadeniz, Bettina Berendt, Sercan Kiyak, Stefan Mertens, and Leen d'Haenens. 2022. Political representation bias in DBpedia and Wikidata as a challenge for downstream processing. *arXiv preprint arXiv:2301.00671* (2022).

[33] Amanda Kau, Xuzeng He, Aishwarya Nambissan, Aland Astudillo, Hui Yin, and Amir Aryani. 2024. Combining knowledge graphs and large language models. *arXiv preprint arXiv:2407.06564* (2024).

[34] Elisavet Koutsiana, Gabriel Maia Rocha Amaral, Neal Reeves, Albert Meroño-Peñuela, and Elena Simperl. 2023. An analysis of discussions in collaborative knowledge engineering through the lens of Wikidata. *Journal of Web Semantics* 78 (2023), 100799.

[35] Qiwei Lin, Derek Ouyang, Cameron Guage, Isabel O Gallegos, Jacob Goldin, and Daniel E Ho. 2025. Enabling disaggregation of Asian American subgroups: a dataset of Wikidata names for disparity estimation. *Scientific Data* 12, 1 (2025), 580.

[36] Vincent Lully, Philippe Laublet, Milan Stankovic, and Filip Radulovic. 2018. Enhancing explanations in recommender systems with knowledge graphs. *Procedia Computer Science* 137 (2018), 211–222.

[37] Yessica Macià, Laura Fernández, and Núria Ferran Ferrer. 2025. Editorial decision-making on Wikipedia: an analysis of gender bias and its impact on discoverability and information retrieval. *Data Technologies and Applications, 2025* (2025).

[38] Maari Maitreyi. 2023. Wikidata: why we contribute to the robot epistemology. https://whoseknowledge.org/wikidata-robot-epistemology/. [Accessed 18-11-2024].

[39] Michael Mandiberg. 2023. Wikipedia's race and ethnicity gap and the Unverifiability of whiteness. *Social Text* 41, 1 (2023), 21–46.

[40] Jennifer Mickel. 2024. Racial/Ethnic Categories in AI and Algorithmic Fairness: Why They Matter and What They Represent. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2484–2494.

[41] Fedor Moiseev, Zhe Dong, Enrique Alfonseca, and Martin Jaggi. 2022. SKILL: Structured knowledge infusion for large language models. *arXiv preprint arXiv:2205.08184* (2022).

[42] Subalalitha Chinnaudayar Navaneethakrishnan, Sathiyaraj Thangasamy, Nithya R, and Neechalkaran. 2022. Exploring the Opportunities and Challenges in Contributing to Tamil Wikimedia. In *International Conference on Speech and Language Technologies for Low-resource Languages*. Springer, 253–262.

[43] Giang Hoang Nguyen, Daniel Beßler, Simon Stelter, Mihai Pomarlan, and Michael Beetz. 2024. Translating universal scene descriptions into knowledge graphs for robotic environment. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 9389–9395.

[44] Tracy Perkins, Sophia Hussein, Mariam Trent, and Lundyn Davis. 2024. Wikipedia and the Outsider Within: Black Feminism and Social Inequality in Knowledge Sharing. *Civic Sociology* 5, 1 (2024).

[45] Pamela Perry. 2001. White means never having to say you're ethnic: White youth and the construction of "cultureless" identities. *Journal of Contemporary Ethnography* 30, 1 (2001), 56–91.

[46] Lydia Pintscher. 2021. Quote from 'Decolonizing the Internet's Structured Data' conference. https://whoseknowledge.org/wp-content/uploads/2022/04/DTI-SD-SummaryReport-ENGLISH.pdf. [Accessed 18-11-2024].

[47] Tim Reeskens and Marc Hooghe. 2010. Beyond the civic–ethnic dichotomy: Investigating the structure of citizenship concepts across thirty-three countries. *Nations and nationalism* 16, 4 (2010), 579–597.

[48] Arjan Reijerse, Kaat Van Acker, Norbert Vanbeselaere, Karen Phalet, and Bart Duriez. 2013. Beyond the ethnic-civic dichotomy: Cultural citizenship as a new way of excluding immigrants. *Political Psychology* 34, 4 (2013), 611–630.

[49] Jonne Sälevä and Constantine Lignos. 2021. Mining Wikidata for Name Resources for African Languages. *arXiv preprint arXiv:2104.00558* (2021).

[50] Salomi Samsudeen, Mohammed Hasan Ali, C Chandru Vignesh, MM Kamruzzaman, Chander Prakash, Tamizharasi Thirugnanam, and J Alfred Daniel. 2023. Context-specific discussion of Airbnb usage knowledge graphs for improving private social systems. *Journal of Combinatorial Optimization* 45, 2 (2023), 66.

[51] Knowledge Graphs Seminar, Nahor Gebretensae, and Heiko Paulheim. 2019. Wikidata: A free collaborative knowledge graph. (2019).

[52] Suman Seth. 2009. Putting knowledge in its place: science, colonialism, and the postcolonial. *Postcolonial studies* 12, 4 (2009), 373–388.

[53] Zaina Shaik, Filip Ilievski, and Fred Morstatter. 2021. Analyzing race and citizenship bias in Wikidata. In *2021 IEEE 18th international conference on mobile Ad Hoc and smart systems (MASS)*. IEEE, 665–666.

[54] Kartik Shenoy, Filip Ilievski, Daniel Garijo, Daniel Schwabe, and Pedro Szekely. 2022. A study of the quality of Wikidata. *Journal of Web Semantics* 72 (2022), 100679.

[55] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. 2020. Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1405–1414.

[56] Valery A Tishkov. 2000. Forget the nation: post-nationalist understanding of nationalism. *Ethnic and Racial studies* 23, 4 (2000), 625–650.

[57] Jan-Philipp Töberg, Axel-Cyrille Ngonga Ngomo, Michael Beetz, and Philipp Cimiano. 2024. Commonsense knowledge in cognitive robotics: a systematic literature review. *Frontiers in Robotics and AI* 11 (2024), 1328934.

[58] Houcemeddine Turki, Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, Lane Rasberry, and Daniel Mietchen. 2023. Ten Years of Wikidata: A Bibliometric Study.. In *Wikidata@ ISWC*.

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

[60] Katy Weathington and Jed R Brubaker. 2023. Queer identities, normative databases: Challenges to capturing queerness on Wikidata. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–26.

[61] Wikimedia. 2021. Wikidata Community Survey 2021. https://upload.wikimedia.org/wikipedia/commons/8/88/Wikidata_Community_Survey_2021.pdf. [Accessed 13-05-2024].

[62] Haotong Yang, Zhouchen Lin, and Muhan Zhang. 2022. Rethinking knowledge graph evaluation under the open-world assumption. *Advances in Neural Information Processing Systems* 35 (2022), 8374–8385.

[63] Charles Chuankai Zhang, 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 (2022), 1–21.

A Property/Entity Descriptions

Property/Entitiy	**Identifier**	**Description and Aliases**
Country of Citizenship	P27	the object is a country that recognises the subject as its citizen *subject of (country)* | *citizenship* | *citizen of* | *nation of citizenship* | *national of* | *(legal) nationality* | *nationality* | *country* | *nation*
Ethnic Group	P172	subject's ethnicity (consensus is that a VERY high standard of proof is needed for this field to be used. In general this means 1) the subject claims it themselves, or 2) it is widely agreed on by scholars, or 3) is fictional and portrayed as such) *ethnicity* | *culture* | *people* | *(cultural) nationality* | *race* | *ethnic or cultural origin* | *tribal affiliation*
Tribe	P12011	recognised membership in a society, mainly denoted by shared cultural heritage; for ethnicity use P172 *member in tribe* | *member of tribe* | *tribal affiliation*
Nationality	Q231002	a legal identification of a person in international law, establishing the person as a subject, a national, of a sovereign state

Table 1. Descriptions and aliases of properties *"Country of Citizenship"*, *"Ethnic Group"*, and *"Tribe"*, and the entity *"Nationality"* found on Wikidata property and entity pages.

B Top 100 Query Results

No.	Country of Citizenship	Count	Ethnic Group	Count	Tribe	Count
1	United States	610763	Han Chinese people	39739	Amazons	50
2	France	366508	African Americans	8437	Ngāpuhi	26
3	Germany	311974	Armenians	2715	Ngāti Kahungunu	15
4	Japan	240953	Minangkabau	2664	Ngāti Porou	14
5	United Kingdom	217908	White Brazilians	2621	Ngāti Awa	13
6	Spain	160098	Sudeten Germans	1798	Igbo	12
7	Italy	146855	Czechs	1595	Ngāi Tahu	12
8	Soviet Union	132249	Jewish people	1464	Ngāi Tūhoe	12
9	Ming dynasty	124300	Pardo Brazilians	1443	Ngāti Tūwharetoa	12
10	Poland	116189	Germans of Hungary	1109	Te Arawa	9
11	Canada	111901	Manchu	1091	Cherusci	9
12	Sweden	110772	Bamar	957	Modoc Nation	8
13	Indonesia	103765	Germans of Romania	744	Ngāti Maniapoto	7
14	Kingdom of the Netherlands	101523	Māori	634	Ngāti Whakaue	7
15	Russia	96322	African Brazilians	599	Ngāti Raukawa	7
16	Brazil	90155	Mongols	588	Te Āti Awa	6
17	India	85440	Odia people	560	Yoruba	6
18	United Kingdom of Great Britain and Ireland	81113	White Americans	525	Ngāti Toa	6
19	Finland	75276	German-Russians	524	Te Rarawa	6
20	Switzerland	69342	Baltic Germans	495	Tainui	5
21	Norway	67120	Americans	469	Bafumbira	5
22	Belgium	66077	Japanese people	459	Waikato Tainui	5
23	Austria	64442	Russians	441	Kāti Māmoe	4
24	Australia	63728	Ukrainians	399	Sac and Fox Nation	4
25	Kingdom of Italy	62924	Yoruba people	378	Banu Tamim	4
26	Hungary	62827	Black people	376	Atrebates	4
27	Qing dynasty	60845	British	364	Ngāti Kahu	4
28	People's Republic of China	54857	Japanese Canadians	359	Te Aitanga-a-Māhaki	4
29	Kingdom of Denmark	54712	Chinese Indonesians	340	Te Aupōuri	4
30	Russian Empire	52543	German minority in Poland	339	Rongowhakaata	4
31	Czechoslovakia	52046	Banat Swabians	318	Waitaha	3
32	Ukraine	43941	White people	313	Ngāti Te Ata	3
33	Song dynasty	42648	Poles	306	Ngāti Pūkenga	3
34	South Korea	42248	English people	305	Ngāti Māhanga	3
35	Slovenia	42143	Transylvanian Saxons	292	Te Māhurehure	3
36	Argentina	41744	Inuit	292	Iceni	3
37	Tang dynasty	37820	Germans	257	Onondaga Nation	3
38	Romania	35950	Navajo	254	Kiga people of Kigezi	3
39	Israel	35772	Greeks	253	Mutayr	3
40	Empire of Japan	34817	Kurds	249	Hunkpapa	3
41	New Zealand	33443	Carpathian Germans	221	Yoruba people	3
42	Mexico	33114	Khitan people	189	Dawasir	3
43	Czech Republic	31304	Germans of Serbia	188	Ngāti Rangiwewehi	3
44	Turkey	30516	Swedish-speaking population of Finland	184	Ngāti Kurī	3
45	Estonia	29761	Arabs	176	Ngāti Paoa	3
46	Greece	29727	Igbo people	174	Whakatōhea	3
47	Taiwan	24618	Northern Sámi people	170	Ngāti Apa	3
48	Portugal	23137	Irish people	169	Ngāti Whātua	3
49	Iran	20562	Vietnamese people	165	Ngāti Ranginui	3
50	German Democratic Republic	19989	Italians	152	Ngāi Tai ki Tāmaki	3

No.	Country of Citizenship	Count	Ethnic Group	Count	Tribe	Count
51	Peru	19682	Caucasian race	145	Ngāti Hine	3
52	Yuan dynasty	19647	American Jews	142	Ngāi Te Rangi	3
53	Uruguay	18182	Mexican Americans	140	Ngātiwai	3
54	South Africa	18151	Italian Americans	137	Quraysh	3
55	Nigeria	18073	Hopi people	136	Baganda	3
56	Ireland	17254	Irish Americans	136	Te Whānau-ā-Apanui	3
57	British Raj	17107	Afro-Uruguayan	136	Ngāti Rongomaiwahine	2
58	Kingdom of Great Britain	17004	First Nations	135	Yuhaaviatam of San Manuel Nation	2
59	Bulgaria	16928	Ashkenazi Jews	135	Ngāti Ruapani	2
60	Polish–Lithuanian Commonwealth	16163	Slovaks	134	Ngāti Tama	2
61	Chile	15366	Belarusians	133	Rangitāne	2
62	Belarus	13221	Cherokee	132	Shammar	2
63	Colombia	13152	Georgians	131	Shawnee	2
64	German Reich	13144	Swedes	127	Ngāti Manawa	2
65	Austria-Hungary	12871	Sámi people	127	Ngāti Hau	2
66	Pakistan	12733	Romani people	126	Ngāti Manu	2
67	Egypt	12246	Malays	125	Ngāti Hako	2
68	Dominion of India	11985	Scottish people	125	Piegan Blackfeet	2
69	Lithuania	11705	Slovenian Germans	122	Te Āti Haunui-a-Pāpārangi	2
70	Kingdom of England	11388	Azerbaijanis	122	Acholi people	2
71	Thailand	11132	Tanguts	120	Bruttii	2
72	Serbia	10558	Hausa people	119	Cananefates	2
73	Cisleithania	10491	Koreans	115	Keliko people	2
74	Socialist Federal Republic of Yugoslavia	9881	Japanese Americans	114	Banu Khazraj	2
75	Latvia	9587	Sioux	114	Bani Khalid	2
76	Slovakia	9480	Thai people	107	Mashpee Wampanoag Tribe	2
77	Ancient Rome	8968	Ojibwe	104	Dakota people	2
78	Croatia	8913	Israeli Jews	104	Navajo Nation	2
79	Philippines	8766	Métis	103	Batavi	2
80	Azerbaijan	8681	Regat Germans	103	Nez Perce	2
81	Ottoman Empire	8510	Filipinos	102	Green Claw	2
82	Kingdom of Yugoslavia	8477	Mediterranean race	99	Rakyat	2
83	Austrian Empire	8306	Norwegians	95	Mohawk	2
84	Venezuela	8275	Palestinians	94	Ngāti Hauiti	1
85	Malaysia	7593	Tatars	92	Mohegan Tribe	1
86	Bangladesh	7422	South Africans	92	Colville tribe	1
87	Cuba	6957	Mon people	88	Native American tribe	1
88	Republic of China	6877	Cree	86	Edoid languages	1
89	Iceland	6761	Rakhine	84	Edo people	1
90	Ghana	6716	Korean Americans	83	Mawsillu	1
91	Kingdom of Prussia	6683	Keliko people	83	Lamtuna	1
92	Armenia	6520	German Americans	80	Southern Ute Indian Tribe	1
93	Algeria	6485	Basque people	78	Santal people	1
94	Morocco	6462	Macedonians	77	Mariche	1
95	Luxembourg	6311	Shan people	77	Atfalati	1
96	Tunisia	6115	Kankalis	75	Al-Assiry	1
97	Kazakhstan	5955	Chinese Americans	73	Lurish tribe	1
98	Kingdom of Portugal	5853	Romanians	69	Ngāti Mahuta	1
99	Saudi Arabia	5696	Chin people	68	Ngāti Maru	1
100	Vietnam	5678	Kiowa people	68	Ngāti Mutunga	1

Table 2. Top 100 SPARQL query results for the properties "Country of Citizenship", "Ethnic Group", and "Tribe"

C Appendix C

```
SELECT ?relatedProperty ?relatedPropertyLabel
WHERE {
    wd:P172 wdt:P1659 ?relatedProperty .
    SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en" . }
}
ORDER BY ?person_label
```

Fig. 7. SPARQL query to acquire related properties to Ethnic Group, used on the Wikidata Query Service.

D Appendix D

```
SELECT ?citizenship ?citizenshipLabel (COUNT(?subject) AS ?count)
WHERE {
    ?subject wdt:P27 ?citizenship .

    ?citizenship rdfs:label ?citizenshipLabel .
    FILTER (lang(?citizenshipLabel) = "en") .
}
GROUP BY ?citizenshipLabel ?citizenship
ORDER BY DESC(?count)
```

Fig. 8. SPARQL query to count the number of people per Country of Citizenship entity. The query was repeated for P172 and P12011.

E Appendix E

```
SELECT DISTINCT ?ethnicGroup ?ethnicGroupLabel ?country ?countryLabel ?continent ?continentLabel
WHERE {
    ?subject wdt:P172 ?ethnicGroup .

    ?ethnicGroup wdt:P17 ?country .
    ?country wdt:P30 ?continent .

    ?ethnicGroup rdfs:label ?ethnicGroupLabel .
    FILTER (lang(?ethnicGroupLabel) = "en") .

    ?country rdfs:label ?countryLabel .
    FILTER (lang(?countryLabel) = "en") .

    ?continent rdfs:label ?continentLabel .
    FILTER (lang(?continentLabel) = "en") .
}
```

Fig. 9. SPARQL query to find associated country (P17) and continent (P30).

F Appendix F

List of Wikidata talk pages that were analysed

- https://www.wikidata.org/wiki/Property_talk:P27
- https://www.wikidata.org/wiki/Property_talk:P172
- https://www.wikidata.org/wiki/Property_talk:P12011
- https://www.wikidata.org/wiki/Wikidata:Property_proposal/Nationality
- https://www.wikidata.org/wiki/Wikidata:Property_proposal/nationality_(cultural_identity)
- https://www.wikidata.org/wiki/Wikidata:Property_proposal/Cultural_identity
- https://www.wikidata.org/wiki/Wikidata:Property_proposal/Tribe
- https://www.wikidata.org/wiki/Wikidata:Property_proposal/Archive/1#P27

We also selected 71 archived discussions, the titles of which are listed below:

- Contemporary constraint 2019/02
- Updating Swedish-speaking Finns in Wikidata with Wikipedia categories as source and reference 2019/11
- Citizenship references 2022/11
- Storing in Wikidata the preferred language variant used on Wikipedia? 2019/08
- Recent addition of languages spoken, written or signed (P1412) on multiple items 2021/04
- How to express nationality (Q231002) ? 2021/04
- Wikidata properties for tribes? 2021/04
- Splitting country of citizenship (P27) 2018/09
- Nationality 2018/09
- Correctly representing archaic or debunked "racial" term entries 2020/01
- Indigenous peoples of the United States 2020/01
- Citizenship status of people before their country existed 2018/02
- Proper way to contribute surnames with a bot 2020/09
- Clean up country / country of citizenship 2019/10
- Singer from Taipei 2019/10
- P31=Norse human 2018/04
- There is no country of citizenship (P27) of Austria-Hungary (Q28513) 2020/10
- Removal of ethnic group without sources 2020/10
- Matching with OpenRefine 2020/06
- Islamic Republic of Afghanistan 2020/06
- How to look up persons nationality? 2017/08
- Dubious citizenships 2018/08
- Property for nationality 2021/11
- Best practices for sourcing values of country of citizenship (P27) 2021/11
- Nationality for ancient people 2021/11
- Germans (Q42884), only for the group or also individuals? 2017/06
- Should persons' former citizenships be deleted? 2018/10
- Denmark 2018/10
- Primary sources tool 2018/10
- UK 1922 or 1927? 202/02
- Complex demographics 2014/02
- Ethnicity 2014/02
- Ethnicity statements again 2022/09
- Fiction countries and country of citizenship 2018/07
- P27 and UK citizenship 2018/07
- Harvest ethnic group Baltic German – estimated 1600 new statements 2018/05
- Edit filter to prevent edits that are obviously wrong? 2019/07
- flaws in property ethnic group (P172) 2017/02
- Ethnic communities and diasporas 2017/02
- removing unreferenced ethnicity rather than finding a reference 2021/12
- Language differences 2021/12
- Former citizenship(s) of people – present or not and if then how? (ideological issue) 2020/04
- Question about 5-year-old vandalism 2020/04

- Wrong nationalities 2016/10
- P131, P172 for languages 2014/07
- Wu Zhen : People's Republic of China painter (1280–1354) ????? 2015/09
- What to do with statement without any source? 2019/09
- Something wrong here? 2020/03
- What's the best way to add the ethnic composition of a region? 2020/08
- Help Trying to query Panama Papers and Citizenship or Place of Birth 2017/09
- Nationality automatic statements 2017/09
- Dutch nationality 2016/07
- Constraints for sources 2016/07
- P of Q 2014/09
- Poorly cited, possibly wrong 2023/04
- Country of citizenship 2014/05
- Removing P27 2022/03
- country of citizenship (P27) on properties 2017/10
- Citizenship 2016/04
- P27 Country of citizenship 2016/11
- Sovereign states, not countries 2021/01
- Help with Q65052377 2019/07
- Useful or not, anti-Semitic or not? 2024/02
- Q56222412
- population of a country 2022/02
- Puerto Rican nationality 2018/09
- "Greenlanders" Q101444758 2020/11
- Are Category:People from Germany (Q7605094) and Category:German people (Q3919754) the same? 2020/04
- Nationality of Jong Tae-se (Q310673) and some bass of Chongryon (Q1058121)
- Question about nationality ?
- People from Brazil 2014/05