Analyzing Race and Citizenship Bias in Wikidata

Zaina Shaik
Information Sciences Institute
University of Southern California
Marina del Rey, CA, USA
zainas@isi.edu

Filip Ilievski
Information Sciences Institute
University of Southern California
Marina del Rey, CA, USA
ilievski@isi.edu

Fred Morstatter
Information Sciences Institute
University of Southern California
Marina del Rey, CA, USA
fredmors@isi.edu

***Abstract*—Since there are limited full-time contributors to Wikidata, the current information might have a bias. In this paper, we examine the race and citizenship bias in general and in regards to STEM representation for scientists, software developers, and engineers. By comparing Wikidata queries to real-world datasets, we discovered that there is an overrepresentation of white individuals and those with citizenship in Europe and North America; the rest are generally underrepresented. We plan to create and implement a bot using a table-linking software to take missing information from the external datasets and insert it into Wikidata to increase minority race representation.**

***Index Terms*—knowledge graph, Wikidata**

I. Introduction

Data is the most powerful when it is accurately represented. A knowledge graph is a collection of knowledge that uses a graph structured data model to represent objects through their attributes and relationships to other objects. Wikidata [1] is an open knowledge graph that contains encyclopedic knowledge, similar to Wikipedia, only in a structured form. As an open and collaborative knowledge graph created by users and bots, it is possible that the data in Wikidata is biased in regards to factors such as gender, race, and country of citizenship.

Indeed, prior work shows that there exists a gender disparity in Wikidata with women being underrepresented [2]. However, by including more data representing women, it is possible to decrease some of the gender bias present [2]. Personalized knowledge graphs are known to have people bias and algorithm bias [3]. Data visualizations about gender biases can help bring awareness to the gender bias in Wikidata.[1] Such biases over Wikidata could be analyzed and visualized by existing tools, such as the Knowledge Graph Toolkit [4] and SPARQL[2].

Biased data leads to inaccurate representation which is harmful because it can influence the perspective of its viewers. Race and country of citizenship representation is necessary for an accurate portrayal of knowledge. Inspired by [2], we tackle the issue of race and country bias in Wikidata by looking at the representation of these groups in Wikidata compared to external datasets.

Furthermore, careers in STEM already have a known lack of representation with non-European countries and ethnicities. We explored whether there was a similar underrepresentation in Wikidata. We found queries for scientists, software developers, and engineers that have an ethnicity and country of citizenship property recorded. After comparison and analysis, we found that for most subcategories of STEM careers, there is an overrepresentation of individuals of the white race and from European and North American countries.

We hope to alleviate the bias by inputting more data from underrepresented groups of races and countries of citizenship into Wikidata. By increasing the representation present in Wikidata, we can help create a better perspective of the backgrounds of individuals with STEM careers and in general.

II. Method

A. Research Questions and Hypotheses

We looked at three research questions. First, how biased is Wikidata in terms of race and country of citizenship representation? We hypothesize that bias exists towards the white race and European countries because only a few Wikidata editors contribute most of the Wikidata edits [5] and there has been evidence of gender bias skewing towards men. Second, how biased is Wikidata in terms of race and citizenship representation for scientists, software developers, and engineers? We hypothesize that bias exists towards the white race and European countries due to existing biases in STEM. Finally, we investigate how we can alleviate the bias by adding more racial and country of citizenship representation with a bot. This tool would let us instill more representation and create a better view of underrepresented communities.

B. Datasets

Wikidata We ran Wikidata queries using SPARQL and KGTK to collect the information already available on Wikidata. Starting by finding datasets to observe racial bias, we looked at the first 50 results of humans with an Ethnicity property (P172) in Wikidata in general. Then, we did the same with scientists, software developers, and engineers. Since Wikidata only contains ethnicity and not race information, we manually grouped the ethnicities into the following race categories: Asian, White, Black, Middle Eastern, Indigenous to America, Hispanic/Latino, and Pacific Islander.

To observe country of citizenship bias, we ran Wikidata queries to find the first 50 results of everyone, scientists, software developers, and engineers with a Country of Citizenship property (P27). We manually grouped the countries into these

[1] http://datakolektiv.org/app/WDCM_BiasesDashboard, accessed on July 9, 2021.

[2] https://www.w3.org/TR/sparql11-query/, accessed on: July 9, 2021.

continent categories: Asia, Europe, Middle East, Africa, North America, South America, and Pacific Islands.

Real-world data In order to have a reference point of data available outside of Wikidata, we found external datasets including the population of races in the world and countries in the world. We found information on Black, Asian, Hispanic/Latinx, and Middle Eastern scientists online on several websites. Additionally, we looked at a dataset involving race diversity in tech careers in the United States. The full list of websites can be found at https://shorturl.at/cuwT4. We plan to scrape them automatically into tables.

C. Bias Measurement

Our goal with the obtained Wikidata queries is to compare the percentages of each race or continent subcategory to the external datasets regarding race and country population in the world. By doing this, we will be able to see how much certain groups are overrepresented or underrepresented. The difference of percentage between the Wikidata query and external dataset will be our unit of measurement for bias.

The external data regarding scientists will be compared to existing information on Wikidata. We will look for how many of the scientists were missing (if any) and what information is currently available about them. The rest of the external datasets will also be compared to data in Wikidata by a bot.

D. Increasing Representativeness of Wikidata

We ran a table linking software[3] to link the Black scientists dataset to Wikidata. This matched up the table columns to existing entities and properties in Wikidata with a 93.75% accuracy rate. For example, Harold Amos was linked to the Wikidata entity with an identifier `Q5659918`. Missing entities, like Earl S. Bell, will be assigned new Wikidata identifiers and described with their corresponding information.

III. Results

Regarding the race of scientists, software developers, and engineers, Wikidata is skewed towards the white race while underrepresenting other races (Table I). While white people make up 17.80% of the world population, in Wikidata, white scientists made up 83.95%, white software developers - 44.08%, and white engineers - 70.74%.

In terms of the country of citizenship of scientists, software developers, and engineers, Wikidata is skewed towards European and North American countries while underrepresenting other continents. Compared to the world population (Europe: 8.80%, North America: 5.45%), Wikidata overrepresents European and North American scientists (71.06% and 15.43%), European and North American software developers (68.09% and 19.88%), and European and North American engineers (66.82% and 19.51%) (Table II).

TABLE I
COMPARISON OF RACE DISTRIBUTIONS BETWEEN WIKIDATA (WD) AND REAL-WORLD DATA.

	Total		Scientists	Software	Engineers
	WD	real	WD	WD	WD
White	**37.63**	**17.80**	**83.95**	**44.08**	**70.74**
Black	18.60	14.83	9.05	16.95	14.72
Asian	39.35	31.14	1.10	20.34	5.97
Middle Eastern	2.37	24.57	5.46	15.25	6.75
Indigenous to America	0.82	3.71	0.00	1.69	0.61
Hispanic/Latino	1.15	n/a	0.44	0.00	1.15
Pacific Islander	0.18	n/a	0.00	1.69	0.00
Other	n/a	7.95	n/a	n/a	n/a

TABLE II
COMPARISON OF COUNTRY OF CITIZENSHIP DISTRIBUTIONS BETWEEN WIKIDATA (WD) AND REAL-WORLD DATA.

	Total		Scientists	Software	Engineers
	WD	real	WD	WD	WD
Asia	18.06	55.26	7.47	5.46	5.36
Europe	**57.10**	**8.80**	**71.06**	**68.09**	**66.82**
Middle East	1.97	5.55	1.30	1.34	1.32
Africa	0.00	11.90	0.00	0.00	0.00
North America	**16.47**	**5.45**	**15.43**	**19.88**	**19.51**
South America	3.62	7.38	2.76	3.46	5.25
Pacific Islands	2.76	5.66	1.98	1.77	1.74

IV. Conclusions and Future Work

There appears to be a bias towards the white race and European and North American countries in reference to STEM careers. This overrepresentation could be due to the racial and country of citizenship backgrounds of Wikidata contributors.

After running the table linking software on the rest of the scientist datasets, we plan to inject the linked tables into Wikidata by implementing a Wikidata bot. With more analysis and the bot, we expect to be able to increase representation and decrease the race and citizenship bias present in Wikidata.

Acknowledgment

We thank the USC Information Sciences Institute NSF REU site: Human Communication in a Connected World.

References

[1] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, 2014.

[2] M. Klein, H. Gupta, V. Rai, P. Konieczny, and H. Zhu, "Monitoring the gender gap with wikidata human gender indicators," in *Proceedings of the 12th International Symposium on Open Collaboration*, 2016, pp. 1–9.

[3] E. J. Gerritse, F. Hasibi, and A. P. de Vries, "Bias in conversational search: The double-edged sword of the personalized knowledge graph," in *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, ser. ICTIR '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 133–136. [Online]. Available: https://doi.org/10.1145/3409256.3409834

[4] F. Ilievski, D. Garijo, H. Chalupsky, N. T. Divvala, Y. Yao, C. Rogers, R. Li, J. Liu, A. Singh, D. Schwabe, and P. Szekely, "Kgtk: A toolkit for large knowledge graph manipulation and analysis," in *The Semantic Web – ISWC 2020*, J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, and L. Kagal, Eds. Cham: Springer International Publishing, 2020, pp. 278–293.

[5] C. Sarasua, A. Checco, G. Demartini, D. Difallah, M. Feldman, and L. Pintscher, "The evolution of power and standard wikidata editors: comparing editing behavior over time to predict lifespan and volume of edits," *Computer Supported Cooperative Work (CSCW)*, vol. 28, no. 5, pp. 843–882, 2019.

[3] https://github.com/usc-isi-i2/table-linker