Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023)

Wiki-based Communities of Interest: Demographics and Outliers

Hiba Arnaout[1], **Simon Razniewski**[1], **Jeff Z. Pan**[2]

[1]Max Planck Institute for Informatics, Saarbrücken, Germany
[2]The University of Edinburgh, Edinburgh, United Kingdom
harnaout@mpi-inf.mpg.de, srazniew@mpi-inf.mpg.de, j.z.pan@ed.ac.uk

Abstract

In this paper, we release data about demographic information and outliers of communities of interest. Identified from Wiki-based sources, mainly Wikidata, the data covers 7.5k communities, e.g., members of the White House Coronavirus Task Force, and 345k subjects, e.g., Deborah Birx. We describe the statistical inference methodology adopted to mine such data. We release subject-centric and group-centric datasets in JSON format, as well as a browsing interface. Finally, we forsee three areas where this dataset can be useful: in social sciences research, it provides a resource for demographic analyses; in web-scale collaborative encyclopedias, it serves as an edit recommender to fill knowledge gaps; and in web search, it offers lists of salient statements about queried subjects for higher user engagement. The dataset can be accessed at: https://doi.org/10.5281/zenodo.7410436.

Introduction

Motivation

A community consists of a group of people who share a commonality such as geography (*Texans*), religion (*Christian*), ethnicity (*Arab*), or a combination (*Arab Texans*). One commonality that is less often discussed are communities of passion or purpose, the so-called *communities of interest* (Fischer 2001). This refers to groups of people who share a profession, practice, or an interest. For instance, members of *White House Coronavirus Task Force* is a community of practitioners in the medical field. Not to be confused with the much broader community of *all* medical practitioners, we focus on contextualized groups of people. In this case, people who were appointed by the *White House* for a specific task. A second example is recipients of *ACM Fellowship*, rather than *all* computer scientists. This allows for fine-grained analyses over various topics, cultures, time frames, etc.

One standard task for understanding communities is identifying their demographic factors. Demographics are statistical information about a community that includes such factors as gender, occupation, linguistic background, nationality, and location (Ashraf 2020). In geo-based communities, for example, identifying demographics can contribute in local policy making or in understanding consumer behavior for national businesses. In communities of interest, it could contribute to identifying under-represented groups or to studying cultural differences between similar communities across countries or continents. For instance, compiling top demographics facilitates the task of finding outliers, i.e., members that have different characteristics than the majority, e.g., *Deborah Birx* is a female while 86% of the *White House Coronavirus Task Force* members are male. These can contribute in studies of under-represented groups in different settings.

Community	Demographics	Outliers
members of White House Coronavirus Task Force	male republican	Deborah Birx (female) Jerome Adams (independent)
members of Indian National Science Academy	male indian	Jörg Hacker (german)
winners of ACM Fellowship	male american computer science	Susan Nycum (female, lawyer) Calvin Gotlieb (canadian)

Table 1: Demographics and outliers of 3 communities.

While extensive research has been conducted on demographics of large geo-based communities (Chambers 2020; Brass 1996) and topic-specific study cases (Poschmann and Goldenstein 2022; Sun and Peng 2021; Zhou et al. 2020), to the best of our knowledge, this is the first work to address communities of interest, releasing data which covers 16 topics from 4 main domains, namely *culture*, *geography*, *history & society*, and *STEM*.

Contributions

We construct a dataset to capture demographics and outliers about communities of interest from Wikidata (Vrandečić and Krötzsch 2014). Table 1 shows three sample communities of interest, including winners of *ACM Fellowship*. While most winners are *male American computer scientists*, two outliers are *Susan Nycum* who is *not* male neither a computer scientist, but a female lawyer, and *Calvin Gotlieb* who is not *American* but *Canadian*. In a nutshell: we collect communities of interest from Wikidata using pre-defined properties [1]

[1]https://www.wikidata.org/wiki/Wikidata:List_of_properties

indicating *interest*, inspired by the idea for collecting peer groups in (Arnaout et al. 2021). For example *position held* indicates people holding the same public office and *member of* indicates people belonging to the same organization, club or group. We then query its topic and domain using the Wiki-topic tool. In the case of *ACM Fellows*, the domain is *STEM* and the topic is *Computing*. Next, we query Wikidata for demographic statements about members of the community, using demography-describing properties in Wikidata, e.g., `gender (P21)`. We compute the most frequent value for each of the factors, e.g., `occupation-computer scientist` (0.93) and `gender-male` (0.80). Finally, we use common factors to identify outliers in the community. These are members whose characteristics do not fully match the community's demography profile. For example, members who are *not* computer scientists. This results in a group-centric dataset. In addition, we construct a subject-centric dataset, by merging for every subject, the list of salient statements across multiple communities they are a member of.

Our datasets can be downloaded from the Zenodo sharing service at: https://doi.org/10.5281/zenodo.7410436. We release the data in an easy to parse JSON format. We also publish a web interface for friendly browsing: https://wikiknowledge.onrender.com/demographics/.

On usage, we believe this dataset can provide useful seeds in social computing problems, by identifying demographic data for a better analysis of communities of interest, and action-warranting under-representations, e.g., gender or ethnicity-bias in certain communities. Downstream applications include comparing similar communities within the same culture/country, e.g., the majority of winners of the *Latin Grammy Award for Best New Artist* are men while *BET Award for Best New Artist* are women (both American musical awards for newcomers). Moreover, the data can give insights into equivalent communities of interest between different cultures, e.g., the typical *Prime Minister of Iran* is *Shia* but the typical *Prime Ministers of Lebanon* is *Sunni*, which are the two main religious branches of *Islam*. One might be intrigued to look more into the historical reasons of these findings or predict future election winners. Beside social sciences, this data can be useful for knowledge curation of collaborative knowledge bases, such as Wikidata, by providing the curators with edit recommendations about missing information, e.g., a missing profession of a member in a community can be derived from its common profession (the typical *Turing Award* recipient is a computer scientist). Finally, producing salient provenance-extended statements about outliers can increase user engagement in web search (Tsurel et al. 2017), e.g., *unlike 560 out of 636 recorded winners of Presidential Medal of Freedom, Stephen Hawking is not an American, but British, recipient.*

Dataset Creation

Identifying Communities

We choose Wikidata as our source of communities of interest. It is a web-scale collaborative knowledge base which covers 97% of Wikipedia articles (called items). For example, *ACM Fellowship* as an article (https://en.wikipedia.org/wiki/ACM_Fellow) and (https://www.wikidata.org/wiki/Q18748039) as an item.

Wikidata, and structured stores of data in general, are a good source for constructing communities of interest, due to their well-canonicalized entities and properties, i.e., using IDs. For instance, one does not have to worry if an *Oscar* is referred to using the word *Oscar*, or *Academy Award*, or any other wording, but can simply query it using its unique ID `Q19020`. Moreover, Wikidata contains additional information about every item, including temporal signals and links to Wikipedia articles, which allows for more sophisticated use cases, e.g., demographics in a certain era. We pick 6 properties indicating interest or profession (`position held - P39`, `award - P166`, `participant in - P1344`, `candidate in election - P3602`, `nominated for - P4353`, `member of - P463`). We instantiate a SPARQL query [2] with a property of interest and one of its objects (community) to collect its members, e.g., `select distinct ?subject where {?subject wdt:P166 wd:Q18748039}` is used to collect *ACM Fellowship* recipients. A list of subjects is returned, including *Thomas Henzinger*, *Susan Nycum*, *Calvin Gotlieb*, etc.

The outcome of this step is 7.5k communities of interest covering 16 topics and 345k subjects. Given a community of interest, the Wiki-topic tool [3] is queried for top-3 topics. An overview is shown in Table 2 [4]. Note that a community can belong to more than 1 topic, e.g., *Presidents of the Senate of Nigeria* is both related to *Politics* under *History & Society* and to *Africa* under *Geography*.

Defining Demographic Factors

Now that we have the communities of interest with their members and topics, we want to identify their most frequent values, given a set of standard demographic factors (Ashraf 2020). We map each of those to equivalent Wikidata properties (see Table 3). For instance, we identify the nationality of a certain member using property `P27`.

Inferring Demographics and Outliers

At this point, we have all the ingredients to start collecting community demographics and outliers. For every community, e.g., `award-ACM Fellow`:

1. From Wikidata, query values for the predefined demography-properties, e.g., `gender`(*Thomas Henzinger*, *Calvin Gotlieb*, ..) = `male`.
2. Compute relative incidence of each factor-value pair, e.g., # male recipients of *ACM Fellowship*/# recipients of *ACM Fellowship*= 673/839 = 0.80
3. Sort by descending order of relative incidence, e.g., `occupation-computer scientist` (0.93), `gender-male` (0.80), `nationality-U.S.` (0.58), etc.
4. Collect outliers as members with demographic data not matching that of the top-k of the community,

[2] https://query.wikidata.org/

[3] https://wiki-topic.toolforge.org/topic

[4] For readability, we only display topics with ≥ 50 communities.

Topic	Domain	# of Communities	Sample Community (***title***, ***recorded members***, ***sample member***)
Biography	Culture	1358	*Winners of Suffrage Science award*, *35*, *Pippa Goldschmidt*
Literature		416	*Winners of EU Prize for Literature*, *96*, *Magdalena Parys*
Media		1071	*Winners of Academy Award for Best Actress*, *79*, *Emma Stone*
Performing Arts		74	*Winners of Special Tony Award*, *21*, *Judy Garland*
Philosophy & Religion		250	*Popes*, *269*, *John Paul II*
Sports		267	*Winners of NBA Coach of the Year Award*, *52*, *George Karl*
Visual Arts		189	*Members of Royal Academy of Arts*, *840*, *William Etty*
Africa	Geography	191	*Presidents of Uganda*, *13*, *Idi Amin*
Americas		927	*Governors of Wisconsin*, *47*, *Jim Doyle*
Asia		947	*Chiefs of the Philippine National Police*, *23*, *Rodolfo Azurin Jr.*
Europe		2737	*Prime Ministers of Poland*, *44*, *Jan Olszewski*
Oceania		247	*Winners of New Zealand Order of Merit*, *148*, *Charles Higham*
Business & Economics	History	148	*Winners of Queen Elizabeth Prize for Engineering*, *11*, *Tim Berners-Lee*
Education		134	*Chancellors of the University of Oxford*, *28*, *Roy Jenkins*
History		252	*Dukes of Normandy*, *26*, *William the Conqueror*
Military & Warfare		386	*United States secretaries of War*, *49*, *William Howard Taft*
Politics & Government		950	*Leaders of the House of Commons*, *80*, *Andrew Lansley*
Society		63	*Winners of Civil Courage Prize*, *36*, *Alexei Navalny*
Biology	STEM	116	*Winners of Darwin Medal*, *67*, *Ernst Haeckel*
Chemistry		63	*Winners of ACS Award in Pure Chemistry*, *92*, *Stuart Schreiber*
Earth & Environment		76	*Winners of Paleontological Society Medal*, *54*, *Alfred Romer*
Mathematics		75	*Winners of SIAM Fellow*, *537*, *Ernst Hairer*
Medicine & Health		98	*Winners of Pollin Prize for Pediatric Research*, *17*, *Basil Hetzel*
Physics		120	*Winners of Nobel Prize in Physics*, *222*, *Albert Einstein*
Space		55	*Winners of NASA Distinguished Service Medal*, *319*, *Frank Borman*

Table 2: Overview of covered topics and sample communities.

e.g., NOT(`gender-male`) applies to *Susan Nycum* and NOT(`nationality-U.S.`) applies to *Calvin Gotlieb*.

Accuracy of Inferred Information. When inferring non-asserted factors for certain members, one unavoidable challenge is the *correctness* of these inferences. Due to the open-world assumption knowledge bases, such as Wikidata postulate, absent statements can be either false (negative), or true but simply absent (missing positive). Present statements in some cases can also be undetectable using exact-match querying due to potential modelling issues. We remedy these using three heuristics: (i) **The partial completeness assumption PCA** (Galárraga et al. 2013; Dong et al. 2014), which asserts that if a subject has *at least one* object for a given property, then there are no other objects beyond those that are in the knowledge base, e.g., if we have at least 1 award for subject `X` then we assume that their list of awards is complete. (ii) **Hierarchical checks**, where we exploit the type system, i.e., class taxonomy, in search for a contradiction of a certain negated factor. For instance, `occupation-Catholic priest` does *not* hold for subject `X`, but `occupation-Latin Catholic priest` does, and `(Latin Catholic priest, subclassOf, Catholic priest)` is a statement in Wikidata. Hence, if `X` is a *Latin Catholic priest*, they are also a *Catholic priest*. (iii) **Semantic similarity checks** to avoid possible synonymous or near-synonymous contradictions, we compute the sentence similarity between a candidate statement and an existing statement for subject `X`. We do so using SBert (Reimers and Gurevych 2019) with 0.6 as a similarity threshold, e.g., similarity (“*teacher*”, “*professor*”) = 0.62, avoiding the inference that someone is a *professor* but not a *teacher* and vice versa. 95% of the eliminated candidates are due to PCA, 2% due to hierarchical checks, and due to 3% semantic similarity checks.

Factor	Wikidata Property
Gender	sex or gender (`P21`)
Sexual orientation	sexual orientation (`P91`)
Occupation	occupation (`P106`)
Political leaning	member of political party (`P102`)
Religion	religion or worldview (`P140`)
Linguistic background	native language (`P103`)
Ethnicity & race	ethnic group (`P172`)
Nationality	country of citizenship (`P27`)
Location	residence (`P551`)

Table 3: Standard demographic factors and their properties.

Dataset Description

We release two datasets [5] on Zenodo [6] and a browsing interface [7].

Group-centric Dataset

This dataset consists of 7530 rows in English language, with a total size of 64MB in JSON format. The fields of a JSON record are:

- Title ID: title of the community using Wikidata IDs.
- Title label: title of the community using equivalent Wikidata labels.
- Number of recorded members: number of subjects in Wikidata that belong to the community.

[5] The data was collected during December 2022.

[6] https://doi.org/10.5281/zenodo.7410436

[7] https://wikiknowledge.onrender.com/demographics/

- Topics: a list of topics describing the community, e.g., `Culture.Media.Music`.
- Demographic factors: a list of top demographics, each consisting of an ID, a label, and a score. The ID describes the factor using Wikidata identifiers, and label describes it using natural language. The score is the relative incidence within the community.
- Outliers:
  - Reason: a statement on why the following members are considered outliers.
  - Score: a numerical value indicating the frequency of this factor in the community.
  - Members: a list of members for which this factor does not hold.

A sample record [8] from the group-centric dataset:

```
{
  "title": "holders of position Lord
      Mayor of Dublin",
  "recorded_members": 91,
  "topics": ["Geography.Northern_Europe"
      ],
  "demographics": [
              "gender-male",
              "occupation-politician"
  ],
  "outliers": [
    {
      "reason": "NOT(male) unlike 81 out
           of 91 recorded members",
      "members": [
              "Catherine Byrne (female)",
              "Emer Costello (female)",
              "Alison Gilliland (female)"
              ]
    },
    {
      "reason": "NOT(politician) unlike
         47 out of 91 recorded members",
      "members": [
              "John D'Arcy (businessperson
                  )",
              "Dermot Lacey (
                  environmentalist)"
              ]
    }
  ]
}
```

Listing 1: JSON record from Group-centric dataset.

Subject-centric Dataset

We rerun the method described in the previous section but with two adjustments: for a given subject, we merge all outlier statements across different communities and rank them by descending order of incidence. Moreover, we extend the list of demography-properties to *all* possible Wikidata properties. This subject-centric dataset consists of 345435 rows in English language, with a total size of 172MB in JSON format.

The fields of a JSON record are:

- Subject ID: the Wikidata ID of the subject.
- Subject label: its equivalent label.
- Statements: a list of salient statements across all communities of interest this individual is a member of:
  - Statement ID: a statement using Wikidata ids.
  - Statement label: a statement using Wikidata labels.
  - Score: relative incidence.

A sample record [9] from the subject-centric dataset:

```
{
  "subject": "Serena Williams",
  "statements": [
    {
      "statement": "NOT(gender-male) but
           (female) unlike 56 out of 68
          recorded winners of L'Equipe
          Champion of Champions.",
      "score": 0.82
    },
      {
      "statement": "NOT(sport-basketball
          ) but (tennis) unlike 4 out of
          8 recorded winners of Best
          Female Athlete ESPY Award.",
      "score": 0.50
    }
  ]
}
```

Listing 2: JSON record from the Subject-centric dataset.

Technical Details. We run our methodol on a CPU cluster with a total of 5376 CPU cores; Hardware: 42x Dell PowerEdge R6525 server; RAM: 16 GB per core; Pre-processing steps [10] include: identifying communities, querying their topics from the Wiki-topic tool, collecting Wikidata statements about members; Running time of demographics and outliers inference process: 4hr8min.

Browsing Interface

For users who wish to sample rows from each topic, we publish a web interface that can be accessed at: https://wikiknowledge.onrender.com/demographics/. A screenshot is shown in Figure 1. It shows that the user selected *Computing* as their topic of interest. It contains 40 communities about *computing*, one of which is *Winners of Turing Award* with top demographic factors as *male computer scientists from the U.S.* Outliers include *women*, e.g., *Barbara Liskov* and non-*Americans*, e.g., *Tony Hoare* who is from the *U.K.* Users can also directly search for communities of their choice.

On top of this feature, the website offers an entity summarization interface, where users can query for their favorite

[8] For readability we omit some of the listed fields.

[9] For readability we omit some of the described fields.

[10] We did not record the running time of pre-processing steps.

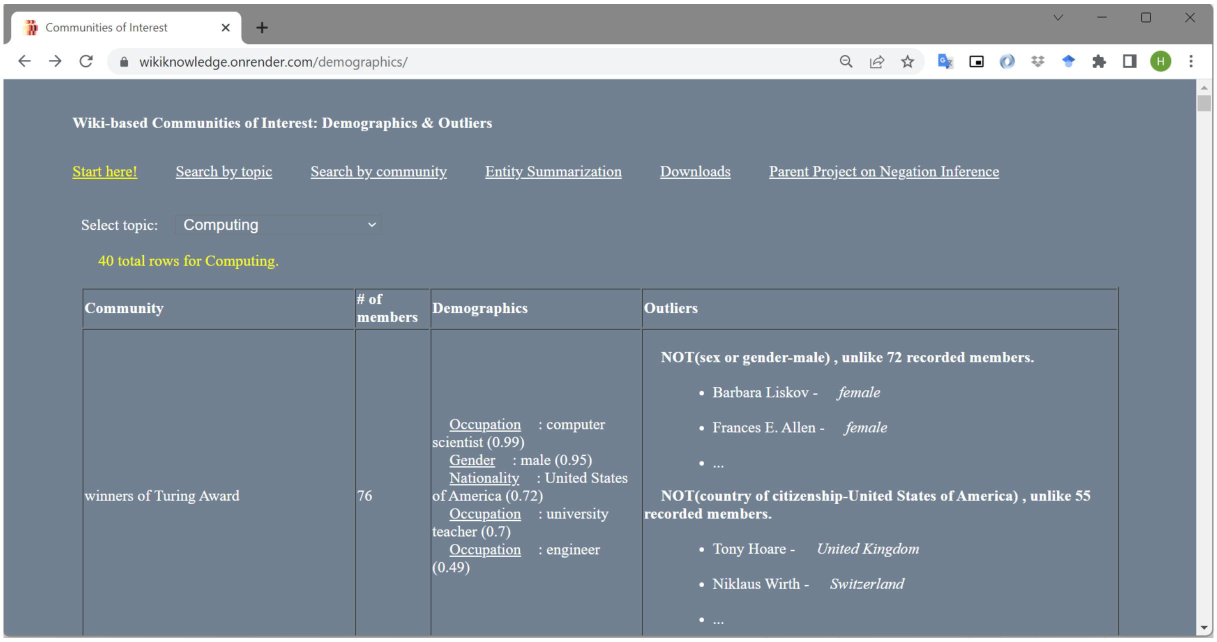

Figure 1: Screenshot of the web interface, searching for demographics and outliers of communities under *Computing*.

public figures and get a list of top salient statements about them, i.e. cases where they were exceptional.

Technical Details. We implement the web portal in Python using Django. For front-end development, we use HTML, CSS, and JavaScript. We deploy it for free on https://render.com.

Applications

Demographic Data Analysis for Social Sciences

Discovering Under-represented Groups. One use case for our data is in academic research in humanities. One standard social problem is identifying under-represented groups (Atkinson et al. 2019; Burek 2021; Clark 1991). These groups can be defined using one or more demographic factor, e.g., *ethnicity*, *gender*. Our data can be considered a resource to answer questions such as: is group X under-represented in community/domain Y? or what is the difference in representation of group X between different domains?

A more specific example is shown in Figure 2. We show the fraction of *female* award recipients (in Wikidata assigned `gender-female`) in STEM [11], and in political offices holders in different continents. These numbers can support needed initiatives for more representation of certain groups, e.g., programs such as *Women in Tech*. In this example, we used Wiki-topics to specify spatial and topical dimensions, e.g., *STEM.Physics* for Physics awards and an intersection of *Regions.Europe and History_and_Society.Politics_and_government* for political offices in certain geographical regions.

Beyond the topics tool, our dataset is not isolated in terms of what we know about communities of interest and their members. Each subject or group can be linked to its Wikidata profile or Wikipedia article, allowing for more customized analyses. For instance, in award winning or holding public offices, statements are normally associated with temporal data in Wikidata, allowing the user of this dataset to explore progress across time. For instance, the charts in Figure 2 can be re-plotted to include time windows, i.e., female award recipients in Physics [1960-1990], Physics [1991-present] and so on.

Exploring Cultural Differences in Governing. Our data can be used in political science research such as understanding governing in different parts of the world. One angle is to understand what kind of professions dominate public offices, e.g., *presidents, mayors, governors, ministers*, etc. We compute these as communities of interest created using the property `position held` [12]. We retain communities of interest under *Politics_and_Government*, and assign each to its equivalent geographical area, e.g., *Central America* (see Table 4).

Note that we drop the profession at rank 1, since it is *politician* for *all* the geographical areas, due to the fact that holding a certain public office automatically turns one into a politician. This data can give better understanding of certain cultures or in case of democracies, how do people vote. *Lawyer* is a recurring profession in Americas, especially in *North America* with quarter of public office holders with `occupation-lawyer`.

[11] https://en.wikipedia.org/wiki/Wikipedia:AfC_sorting/STEM

[12] https://www.wikidata.org/wiki/Property:P39

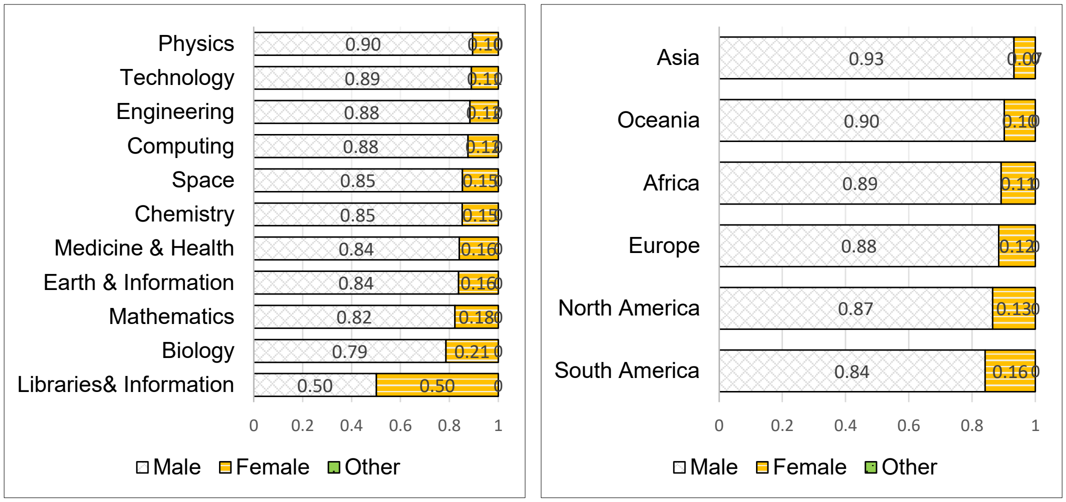

Figure 2: Female award recipients in STEM (left) and female political office holders in different continents (right).

Edit Recommendations for Collaborative Encyclopedias

The Web-scale collaborative knowledge base Wikidata contains more than 100 million items (or subjects) which have received almost 2 billion edits since its inception. Editors often need to prioritize their efforts, so useful tools to guide them can improve data quality and completeness, e.g., the Recoin plugin (Balaraman, Razniewski, and Nutt 2018) helps focus the editing on missing properties of subjects. Our approach and dataset can be used to improve this service by, not only proposing relevant missing properties, but also proposing a full statement about that property. As mentioned, we consider the PCA prior to inferring the negativity of a certain demographic factor. In that step, one cannot assert absent information but can offer a *calculated guess* of what that might be, leaving it for human curators to confirm or deny. For example, *Maja Vuković* is a member of winners of *IBM fellowship*. For the property occupation in Wikidata, she has zero values [13] and Recoin lists `occupation` as the top missing property. Given the demographic data we have about professions of this community she is member of, we propose *computer scientist, mathematician,* and *engineer* as top 3 candidates. This can especially contribute to the completeness of information about long tail entities. Moreover, for subjects who are members of multiple communities, one can merge similar demographic values across communities, then average their confidence scores.

Entity Summarization

Web search results of queries about public figures can be improved by including salient and sometimes surprising facts. It increases user engagement (Tsurel et al. 2017) to augment question answering and entity summarization results with *did you know*-like statements. This is where our second dataset, i.e., the subject-centric, shines. A user can use the autocompletion field in our web interface to query statements about subject of their choice. These are often surprising or unexpected statements. Examples:

- Did you know that unlike 96% of *Oscar for Best Director winners*, *Bong Joon-ho* does not speak *English*, but *Korean*? [14]
- Did you know that unlike 88% of winners of the *Presidential Medal of Freedom*, *Stephen Hawking* is a not an *American*, but a *British*, winner?
- Did you know that unlike 91% of recipients of *Liebig Medal* (established by Association of German Chemists), recipient *Max Planck* was not a chemist but a physicist?

[13] https://www.wikidata.org/wiki/Q111536437

Ethical & FAIR Considerations

In this work, datasets released are based on public information and tools provided by Wikimedia projects. We do not use any personal information. Every dataset record, i.e., JSON object, includes all equivalent labels and IDs of properties and items, retrieved from Wikidata (December 2022 version). Users can refer to these IDs for more details and definitions.

Conclusion

In this work, we generate demographics from Wiki-based sources by exploiting information about communities of interest. We release two datasets and publish a web interface for friendly browsing. Finally, we show three purposes for the data, namely as a resource for social sciences problems, edit recommender for collaborative knowledge bases, and salient fact generator for a better search experience.

Acknowledgments

Funded by the German Research Foundation (DFG: Deutsche Forschungsgemeinschaft) - Project 453095897.

[14] In fact he was accompanied by a translator to deliver his acceptance speech.

Area	Top professions
Central Africa	diplomat (0.27), economist (0.04), civil servant (0.01), philosopher (0.01), minister (0.01)
Eastern Africa	diplomat (0.09), judge (0.03), lawyer (0.03), military personnel (0.03), economist (0.02)
Northern Africa	diplomat (0.12), ruler (0.12), lawyer (0.03), military personnel (0.01), imam (0.01)
Southern Africa	judge (0.28), lawyer (0.11), civil servant (0.01), businessperson (0.01)
Western Africa	diplomat (0.17), lawyer (0.03), military personnel (0.03), economist (0.01), judge (0.01)
Central America	lawyer (0.07), diplomat (0.07), writer (0.02), economist (0.01), military personnel (0.01)
North America	lawyer (0.25), diplomat (0.06), judge (0.03), military personnel (0.01), businessperson (0.01)
South America	lawyer (0.17), diplomat (0.05), military personnel (0.02), journalist (0.01), historian (0.01)
East Asia	monarch (0.09), diplomat (0.07), lawyer (0.06), judge (0.06), prosecutor (0.01)
South Asia	diplomat (0.05), lawyer (0.03), economist (0.02), civil servant (0.02), judge (0.01)
Southeast Asia	sovereign (0.09), judge (0.08), lawyer (0.07), military personnel (0.03), diplomat (0.02)
West Asia	diplomat (0.12), sovereign (0.08), military personnel (0.05), physician (0.02), poet (0.01)
Eastern Europe	diplomat (0.12), economist (0.04), lawyer (0.02), monarch (0.02), university teacher (0.01)
Northern Europe	judge (0.08), diplomat (0.04), monarch (0.02), lawyer (0.02), journalist (0.01)
Southern Europe	diplomat (0.07), lawyer (0.04), military personnel (0.02), jurist (0.01), monarch (0.01)
Western Europe	lawyer (0.13), judge (0.06), diplomat (0.03), military personnel (0.02), suffragist (0.02), teacher (0.01)
Oceania	lawyer (0.08), diplomat (0.04), judge (0.01), pastoralist (0.01), solicitor (0.01), farmer (0.01)

Table 4: Top professions in political offices in different parts of the world.

References

Arnaout, H.; Razniewski, S.; Weikum, G.; and Pan, J. Z. 2021. Negative Statements Considered Useful. *JWS*.

Ashraf, M. A. 2020. Demographic factors, compensation, job satisfaction and organizational commitment in private university: an analysis using SEM. *JGR*.

Atkinson, R.; Lu, P.; Cho, N. L.; Melnitchouk, N.; and Kuo, L. E. 2019. Gender disparities in award recipients from surgical specialty societies. *Surgery*.

Balaraman, V.; Razniewski, S.; and Nutt, W. 2018. Recoin: Relative Completeness in Wikidata. In *WWW Companion*.

Brass, W. 1996. Demographic Data Analysis in Less Developed Countries: 1946–1996. *Population Studies*.

Burek, C. V. 2021. Female medal and fund recipients of the Geological Society of London: a historical perspective. *Geological Society, London, Special Publications*.

Chambers, J. 2020. Hybrid gridded demographic data for the world, 1950-2020. 10.5281/zenodo.3768003.

Clark, J. 1991. Getting There: Women in Political Office. *AAPSS*.

Dong, X. L.; Gabrilovich, E.; Heitz, G.; Horn, W.; Murphy, K.; Sun, S.; and Zhang, W. 2014. From Data Fusion to Knowledge Fusion. VLDB.

Fischer, G. 2001. Communities of interest: Learning through the interaction of multiple knowledge systems. In *IRIS*.

Galárraga, L. A.; Teflioudi, C.; Hose, K.; and Suchanek, F. 2013. AMIE: association rule mining under incomplete evidence in ontological knowledge bases. WWW.

Poschmann, P.; and Goldenstein, J. 2022. Disambiguating and Specifying Social Actors in Big Data: Using Wikipedia as a Data Source for Demographic Information. *Sociological Methods & Research*.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP.

Sun, J.; and Peng, N. 2021. Men are elected, women are married: Events gender bias on wikipedia. *arXiv preprint arXiv:2106.01601*.

Tsurel, D.; Pelleg, D.; Guy, I.; and Shahaf, D. 2017. Fun Facts: Automatic Trivia Fact Extraction from Wikipedia. WSDM.

Vrandečić, D.; and Krötzsch, M. 2014. Wikidata: a Free Collaborative Knowledge base. *CACM*.

Zhou, S.-J.; Zhang, L.-G.; Wang, L.-L.; Guo, Z.-C.; Wang, J.-Q.; Chen, J.-C.; Liu, M.; Chen, X.; and Chen, J.-X. 2020. Prevalence and socio-demographic correlates of psychological health problems in Chinese adolescents during the outbreak of COVID-19. *European child & adolescent psychiatry*.