Semantic Web 1 (0) 1–4
IOS Press

How can the social sciences benefit from knowledge graphs? A case study on using Wikidata and Wikipedia to examine the world's billionaires

Daria Tisch [a,*] and Franziska Pradel [b]
[a] *Max Planck Institute for the Study of Societies*
E-mail: tisch@mpifg.de
[b] *Technical University of Munich*

Abstract. This study examines the potentials of Wikidata and Wikipedia as knowledge graphs for the social sciences. The study demonstrates how social science research may benefit from these knowledge bases by examining what we can learn from Wikidata and Wikipedia about global billionaires (2010-2022). First, knowledge graphs provide human knowledge, which can be used to generate datasets informing about, for example, political, economic, and cultural elites or other notable people. Second, knowledge graphs provide linked (open) data that can be used to examine social networks of a different kind but also enable social scientists to connect different databases to enrich their research data. We show that the English Wikipedia and, to a lesser extent, Wikidata exhibit gender and nationality biased in the coverage and information about global billionaires. Using the genealogical information that Wikidata provides, we examine the family webs of billionaires and show that at least 15% of all billionaires have a family member also being a billionaire. We discuss the challenges and limitations of using Wikidata and Wikipedia for research purposes.

Keywords: Wikidata, knowledge graph, super-rich, Wikipedia, SPARQL, elites

1. Introduction

People seeking knowledge on a particular topic often consult Wikipedia. Wikipedia articles mainly consist of unstructured data written by a community of volunteers in open collaboration. In contrast, although also collaboratively edited, Wikidata consists of structured data, enabling not only humans but also machines to gather and process knowledge efficiently. Whereas Wikipedia was launched already in 2001, its little sister Wikidata was launched in 2012 and has been one of the most often edited knowledge bases ever since [8]. Both knowledge bases are not only an enrichment of individuals' opportunities to gain knowledge, educate themselves, and contribute to the public's knowledge but also a novel data source for research in the social sciences. For example, research on political, cultural, and economic elites often complains about the struggle to find (comparative) data on elites [12, 15]. However, not only elite research but different sub-disciplines within the social sciences may take advantage of the opportunities these knowledge bases offer.

*Corresponding author. E-mail: tisch@mpifg.de.

This study aims at promoting the use of Wikidata and Wikipedia in the social sciences. The study first summarizes the potentials and strengths as well as the challenges and caveats of using such knowledge bases as a social scientist. It then demonstrates the potential of knowledge bases for social science research by focusing exemplary on one area, namely notable individuals. In particular, it is examined what we can learn from Wikidata and Wikipedia about global billionaires. For this purpose, a data set containing US dollar billionaires based on the *Forbes* billionaire lists from 2010 until 2022 is linked to Wikidata (N = 4032). Additional data are scraped using both Wikidata's and Wikipedias' Application Programming Interfaces (API) and analyzed. The study concludes by discussing further applications in the social sciences.

This study's main contribution is two-fold. First, it contributes to the fast-growing computational science literature by showing the benefits of using Wikidata and Wikipedia as research data for the social sciences. Second, this study contributes to the literature on economic elites. It shows that the amount of public information about individual billionaires is linked to their characteristics, such as gender and nationality. Public knowledge about global billionaires is biased toward men and billionaires from North America. It further shows that billionaires are interconnected via family lines hinting at the vital role of the family for billionaires' wealth but also at social closure at the top of the wealth distribution by way of marriages [18, 26].

2. Strengths of Wikidata and Wikipedia for the social sciences

2.1. Wikidata

> Wikidata is a free, collaborative, multilingual, secondary database, collecting structured data to provide support for Wikipedia, Wikimedia Commons, the other wikis of the Wikimedia movement, and to anyone in the world. [29]

But how can social scientists benefit from Wikidata? Let us look at the characteristics of Wikidata to understand Wikidata's strengths for the social sciences. First, it provides structured, human- and machine-readable information on different topics. Social scientists can easily create data sets on different topics from structured data using the query language *SPARQL*. Thus, Wikidata can function as a content provider [36]. For example, scholars studying the elite could examine the diversity within political, economic, and cultural elites by examining socio-demographic characteristics such as gender, age, or education, all properties included in Wikidata. Furthermore, prior research has already used Wikidata to examine the kinship networks of political elites [32].

Second, Wikidata provides linked open data. Thus, it provides not only links within the Wikidata knowledge graph but also links to other databases. For this purpose, Wikidata items include external identifiers. For example, Wikidata entries about actresses can be easily linked to their entries in movie databases. Or Wikidata entries of scientists can be easily linked to their entries in the ORCID database [23]. Individuals on Wikidata can also be linked to their Twitter accounts or entries in different encyclopedias.

Third, Wikidata is collaboratively edited, constantly updated, and improved by a large community of contributors. Thus, it is a growing knowledge base. Because social scientists often generate their own databases, they could not only use already available information but also contribute to Wikidata by importing their data or linking their data to existing datasets [1, 4, 36]. For example, the Comparative Legislators Database, covering over 45,000 contemporary and historical politicians from ten countries, is based on Wikidata and is connected to other political databases via external identifiers [12]. Another example is the *Everypolitician* project that collected rich structured data with sociodemographic and political information about 78,382 politicians from more than 233 countries, also by complementing their data with Wikidata and made the datasets freely available to the public.[1]

Fourth, Wikidata is multilingual, i.e., it supports data in many languages. This makes it easier to connect information across different language versions of Wikipedia, as well as other projects. Wikidata's multilingualism

[1] As of today, extensive data has moved to Wikipedia after the end of the project, and the project initiated a Wikiproject, concurrently with asking the science and wiki community for help to transfer data to Wikipedia and contribute to the data (for more information, see also https://www.wikidata.org/wiki/Wikidata:WikiProject_every_politician and for the Exerypolitican dataset http://everypolitician.org/).

allows social scientists to conduct research without language barriers. For example, elite researchers might benefit from Wikidata's multilingual features because they can easily incorporate members of the global elite from different countries. Wikidata can also be used for named entity linking [9].

Finally, Wikidata is free and transparent. The information stored in Wikidata is available for anyone to use, reuse, and redistribute. Social scientists using open data or making data open by contributing to Wikidata promote transparency and contribute to open science. Wikidata has been utilized as a data hub in different disciplines such as linguistics, digital humanities, medicine, or biology [1, 8, 36]. However, in the social sciences, Wikidata is still rarely used.

2.2. Wikipedia

While many social scientists use social media data, such as from Twitter, far fewer studies use Wikipedia data to answer research questions. As outlined above, Wikidata contains structured data with meta information about persons or topics. This can be distinguished from less structured information from Wikipedia like Wikipedia articles, talk pages, and edit histories.

Wikipedia articles themselves, as well as the associated talk pages, can give insight into controversies and societal discussions. The edit history can uncover interesting social science questions; for instance, one study used this information to investigate editing histories of German political elites' Wikipedia biographies [13], which suggested that they were regularly edited, also from the German Bundestag. Wikipedia enables researchers access to such information that can be used for enriching data sets about political and economic elites - that might otherwise not or only hardly be accessible.

Several studies used computational (social) science methods to investigate Wikipedia articles' content and structured metainformation to investigate social science research questions. Some of these studies focused on examining systemic biases on Wikipedia. In examining Wikipedia data, studies have repeatedly found gender biases in structural terms, in the representation of notable persons in their content in Wikipedia biographies, their visual presentation and metadata [3, 16, 27, 28]. A robust finding is that Wikipedia suffers from a gender bias in biographies of notable persons. Besides structural differences in metadata and hyperlinks, biographies of women contain more stereotypical female topics like information about their relationship status or family status, as well as more negative information than biographies about notable men [28]. Another study found a gender bias in the ratio of words related to a political role to personal information in Wikipedia articles for political elites from conservative parties in Germany [22].

Besides studies on systemic biases on Wikipedia, there is a wide range of applications of Wikipedia data for answering social science questions. For instance, a study used Wikipedia data on oil price-related article views to forecast the crude oil price along with other online media sources and found them to have high predictive power for oil price movement [7]. The number of Wikipedia pages that link to Wikipedia biographies and the number of Wikipedia language versions have been used as a measure of status or importance of the respective person [11, 20, 35], topics or historical events in a society [10]. For example, one study used Wikipedia lists of suicides of notable persons to first identify notable people who committed suicide. It then used page links to the persons' biography as a measure of status to estimate the relationship between celebrity suicides and the general population's suicide rate [20]. Different studies showed the possibility of using Wikipedia to systematically collect an extensive data set on a specific topic by means of categories and lists within Wikipedia (e.g., like in the aforementioned study "List of suicides", or "Lists of members of parliament").

Research also used Wikipedia articles' page views to complement existing data sets with a measure of popularity or public interest for successful persons or the political elite [22, 25] to examine whether biases in media coverage [25] or Google search autocompletion [22] about these elites change when controlling for such measure. Similarly, other studies used page views as a measure of public attention to issues. For instance, a study examined whether celebrity advocacy increases public attention to political issues measured using pageviews [2].

Moreover, different language editions enable social scientists to further investigate research questions across different language editions, for instance, for identifying similarities between different languages as a proxy for cultures [19]. One example is a study that examined similarities in food cultures using Wikipedia articles about food in different Wikipedia language editions and also validated their findings with survey research.

2.3. Application

The data is easily and freely available through the Wikimedia APIs and the Wikidata Query Service (https://query.wikidata.org/). A great online tutorial by Cohen, Baumann, and Munzert gives a hands-on introduction on how to work with Wikidata and Wikipedia as a (comparative) political scientist [6]. It is worth mentioning that there are already different R packages that help scientists to scrape data from Wikidata and Wikipedia easily. The API client library *WikipediR* [14], a wrapper for the MediaWiki API, is useful to easily retrieve metadata of Wikipedia pages such as backlinks to that page, internal wikilinks from that page, external links, or the length of Wikipedia articles measured in bytes. *WikidataR* is a read-write API client library for Wikidata [24]. It is useful to send SPARQL queries to the Wikidata Query Service SPARQL endpoint. If one does not want to write SPARQL queries, this package also helps to retrieve the data associated with individual Wikidata items and properties. It can also be used to write statements to Wikidata. Another helpful R package is *pageviews*, an API Client for Wikimedia Traffic Data. This package enables researchers to access daily, monthly, or yearly page views for Wikipedia articles. The R package *rvest* allows scraping information from Wikipedia, like information stored in tables, easily. To learn writing SPARQL queries, Wikidata's tutorial page (https://www.wikidata.org/wiki/Wikidata:SPARQL_tutorial/en) and the examples at the Wikidata Query Service are very helpful.

3. Caveats of Wikidata and Wikipedia for the social sciences

Although Wikidata and Wikipedia offer many benefits for social scientists, several potential drawbacks and limitations must be considered when using these knowledge bases for research purposes. They are maintained and constantly expanded by a large number of mostly anonymous volunteers. Therefore, the accuracy of the data is not guaranteed, and the perspectives and biases of the editors might influence the information provided. For instance, ideological biases may be persistent in Wikipedia articles, particularly when covering a controversial topic (e.g., about politics, history, religion) [17, 31, 34]. Wikipedia editors are a homogeneous group of more than 90 percent men, most of whom live in North America and Europe and are predominantly editing in English. A 2011 Wikipedia Editor survey indicates that more than 70 percent of edits are contributed to the English language edition. Thus, researchers working with different language editions or on national topics or notable persons for comparative purposes may need to consider such biases.

A related problem concerns the amount of missing information. Missing information on, for example, members or characteristics of the elites might exhibit different kinds of biases such as gender, language, and race bias [27]. Moreover, limited accessibility of Wikipedia due to censorship of whole Wikipedia or selected Wikipedia pages[2] - or due to limited internet access - warrant attention to social science researchers who use Wikidata and Wikipedia data [5, 21, 30, 33]. Limited access to the Wikipedia project may affect editing, coverage of topics as well as measures like article pageviews. By complementing the knowledge bases with other databases, social scientists might help to fill in the missing information.

4. The world's billionaires on Wikidata and Wikipedia

Below, some of the core strengths of knowledge bases and graphs for social science research are highlighted by examining the following research questions: What do Wikidata and Wikipedia know about the world's billionaires? What predicts public knowledge about billionaires? How connected are the world's billionaires via family lines? The first set of questions regards the coverage of billionaires in Wikidata and Wikipedia, and the second question uses Wikidata as a content provider.

[2]Wikipedia gives more information on censorship on their website "Widespread censorship of Wikipedia has occurred in countries including (but not limited to) China, Iran, Myanmar, Pakistan, Russia, Saudi Arabia, Syria, Tunisia, Turkey, Uzbekistan, and Venezuela. Some instances are examples of widespread internet censorship in general that includes Wikipedia content. Others are indicative of measures to prevent the viewing of specific content deemed offensive." (for more information, see https://en.wikipedia.org/wiki/Censorship_of_Wikipedia).

Because Wikipedia is one of the most popular encyclopedias, information about billionaires provided by Wikidata and Wikipedia can be perceived as a proxy for public knowledge about the billionaires. Thus, Wikipedia and Wikidata (besides the *Forbes* website) can be regarded as key platforms for public knowledge about the world's billionaires.

4.1. Data: Forbes billionaires, 2010-2022

The *Forbes* magazine publishes a list of US dollar billionaires every year. Based on the list from 2010 until 2022, a data set including all listed billionaires is generated. The final data set includes a unique identifier, name, gender, birth date, country of citizenship, highest rank, and maximum wealth (between 2010 and 2022) of 4032 billionaires. The 4032 billionaires were reconciled with Wikidata with the help of *OpenRefine*. *OpenRefine* is an open-source tool for working with messy data and includes a reconciliation function. Reconciliation means that string values are matched to entities in Wikidata or other databases. The mapping of the billionaires' names to items in Wikidata was checked manually for all 4032 individuals.

4.2. Examining coverage: How much information do Wikipedia and Wikidata provide about global billionaires?

Figure 1 shows how the coverage of billionaires in Wikidata varies between countries of origin. The coverage is especially low in China, whereas it is particularly high in Norway and Ireland. Table 1 shows summary statistics. 66% of the 4032 billionaires could be matched to a Wikidata item. On average, a billionaire's Wikidata item contains 24 properties. Billionaires have Wikipedia pages in 7 languages, on average, and 57% have an English Wikipedia page. The mean length of Wikipedia pages is 814 words. On average, billionaires have 488 views daily. The English Wikipedia articles about billionaires contain, on average, 81 links to other pages, and 77 Wikipedia pages refer to the respective articles about billionaires.

Table 1
Descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Max
Highest rank	4,032	989.35	674.39	1	2,674
Maximum wealth (in millions)	4,032	4,706.30	9,506.51	1,000	219,000
Number of properties	2,679	23.59	27.39	0	463
Number of Wikipedias (languages)	2,679	6.89	13.50	0	232
Number links to other pages	2,295	80.63	112.91	1	500
Number links to billionaire's page	2,295	77.27	121.92	1	500
Number of words	2,295	813.77	1,291.95	5	18,350
Views	2,295	487.77	6,073.46	1.00	278,888.20
Wikidata (yes-no)	4,032	0.66	0.47	0	1
English Wikipedia article (yes-no)	4,032	0.57	0.50	0	1

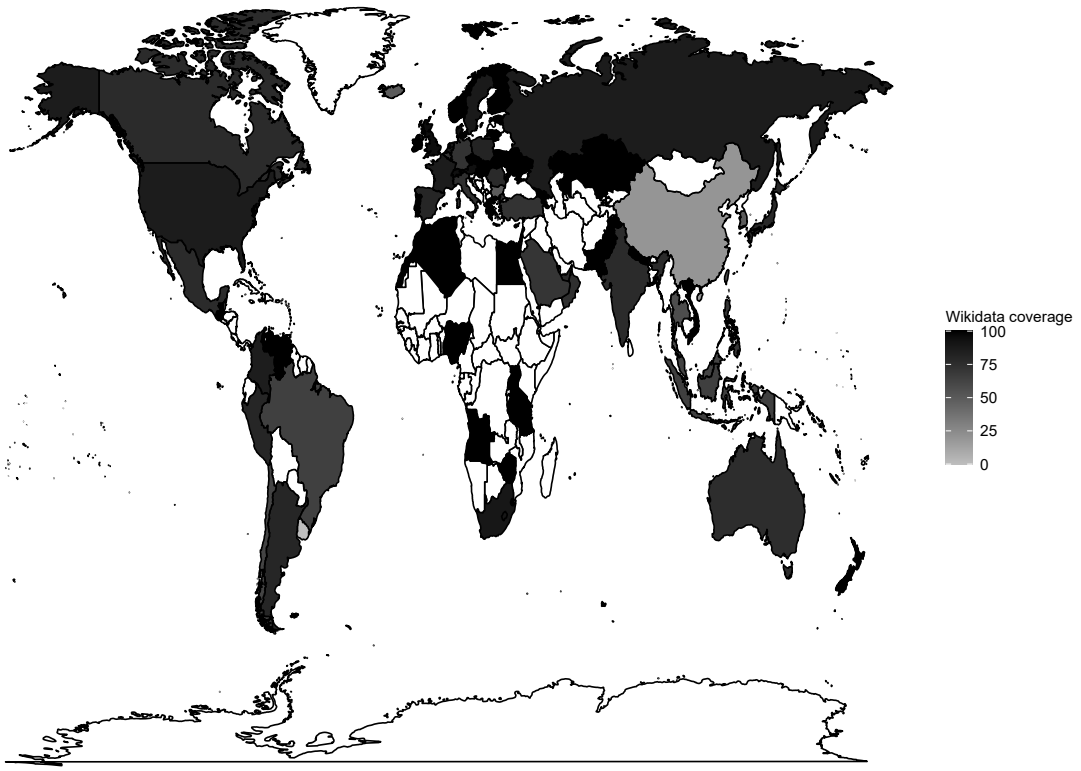

Fig. 1. Coverage of billionaires in Wikidata

4.3. Analysis of bias: Are Wikipedia articles and Wikidata entries about billionaires biased?

4.3.1. Regressions

To assess the data quality, it is important to understand if billionaires with specific characteristics are systematically underrepresented in Wikidata and Wikipedia. Therefore, logistic regressions are used to estimate the probability of being part of Wikidata and the English Wikipedia.

However, not only under-representation of specific groups is problematic, but also the amount of information provided or the kind of information provided is biased. To examine if the amount of information provided by Wikipedia and Wikidata is biased, we estimated ordinary least squares regressions. As dependent variables, we use the number of properties in Wikidata as well as the number of words on the English Wikipedia page.

Figure 2 shows the predicted probabilities of having a Wikidata item (first column) and having an English Wikipedia article (second column) for different groups. For the full regression results, see Table 2. Panels A1 and A2 show that male billionaires are both more likely to have a Wikidata item and to have an English Wikipedia article. Panels B1 and B2 show that the billionaire's year of birth is related to the probability of having a Wikidata item and Wikipedia article. The younger a billionaire, the lower the probability. Last, Panels C1 and C2 show that billionaires from Asia and South America have a lower probability of having a Wikidata item or an English Wikipedia page than billionaires from other continents. To conclude, the logistic regressions hint at gender, age, and nationality bias in the coverage of billionaires.

To examine if also the amount of information about the billionaires provided in Wikidata and the English Wikipedia exhibits biases, Table 3 shows the results of OLS regressions. Regarding the number of properties in Wikidata, female and male billionaires do not seem to differ significantly. The age of billionaires is also not statistically significantly related to the number of properties. However, the amount of a billionaire's wealth is positively

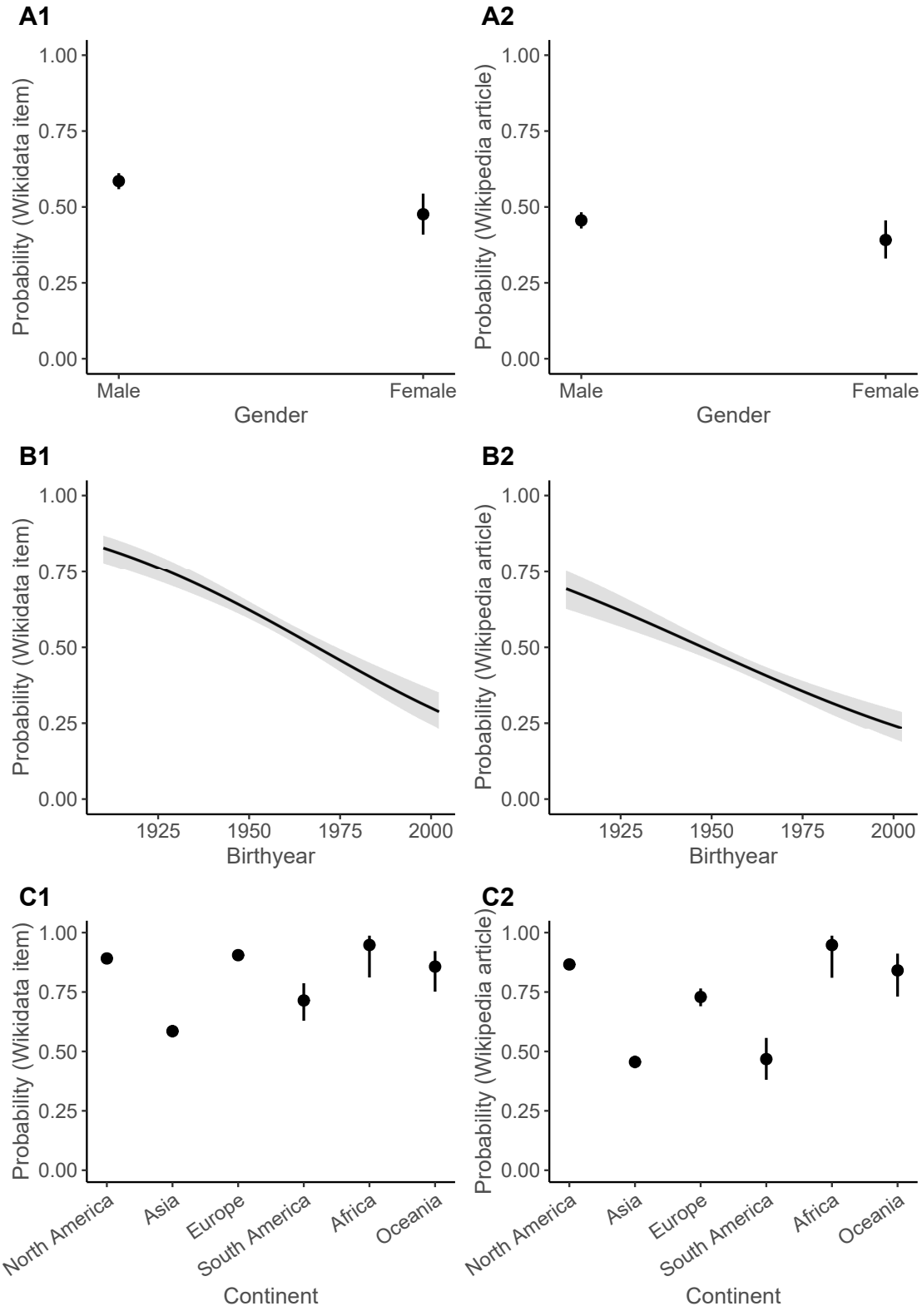

Fig. 2. Predicted probabilities of having Wikidata item or English Wikipedia article

Table 2
Logistic regression on having Wikidata page or English Wikipedia page

	Dependent variable:	
	Wikidata	English Wikipedia
Male	0.440***	0.265**
	(0.137)	(0.132)
Birthyear	−0.027***	−0.022***
	(0.003)	(0.003)
Africa (Ref. North America)	0.797	1.029
	(0.742)	(0.743)
Asia (Ref. North America)	−1.759***	−2.044***
	(0.110)	(0.107)
Europe (Ref. North America)	0.152	−0.876***
	(0.153)	(0.130)
Oceania (Ref. North America)	−0.314	−0.201
	(0.359)	(0.350)
South America (Ref. North America)	−1.186***	−1.995***
	(0.219)	(0.204)
Rank	−0.001***	−0.001***
	(0.0001)	(0.0001)
Intercept	55.270***	45.052***
	(6.140)	(5.735)
Observations	3,851	3,851
Log Likelihood	−1,776.760	−1,896.944
Note:		*p<0.1; **p<0.05; ***p<0.01

related to the number of properties in Wikidata (a higher rank value indicates less wealth). Wikidata also exhibits a nationality bias. On average, the number of properties is higher for North American billionaires than for any other billionaire.

Biases are more pronounced for the length of the English Wikipedia article measured in the number of words. Ceteris paribus, articles about male billionaires are estimated to have 259 more words than articles about female billionaires. The younger a billionaire, the more words their article contains on average. In addition, articles about North American billionaires are, on average, longer than articles about any other billionaire. However, because only English Wikipedia articles are studied, it is not surprising that articles about North Americans are longer. Last, again, the regression shows that the amount of wealth relates significantly to the number of words.

Table 3
OLS regressions on number of properties on Wikidata page and number of words in English Wikipedia page

	Dependent variable:	
	Number of properties	Number of words
Male	0.362	258.688***
	(1.737)	(87.161)
Birthyear	0.043	3.804**
	(0.036)	(1.818)
Africa (Ref. North America)	−2.788	−459.623**
	(4.422)	(208.035)
Asia (Ref. North America)	−9.677***	−553.587***
	(1.285)	(64.747)
Europe (Ref. North America)	−3.764***	−474.841***
	(1.417)	(71.751)
Oceania (Ref. North America)	−7.418*	−110.045
	(4.002)	(190.223)
South America (Ref. North America)	−11.167***	−675.392***
	(2.775)	(152.233)
Rank	−0.008***	−0.247***
	(0.001)	(0.047)
Intercept	−48.718	−6,350.482*
	(70.554)	(3,546.278)
Observations	2,600	2,242
R^2	0.052	0.052
Note:		*p<0.1; **p<0.05; ***p<0.01

4.3.2. Scatter plots

To get more explorative insights into the relationship between Billionaires' online visibility and representation, we visualized the mined average daily page views and the number of words in the Wikipedia articles along with billionaires' wealth.

The three scatter plots in Figure 3 show the relationship between the i) wealth of billionaires, the ii) Wikipedia page views, and iii) article length. The first plot (A) in the upper panel shows a positive association between the average number of page views and the number of words in billionaires' Wikipedia biographies, demonstrating that more popular billionaires often have longer Wikipedia biographies. In the lower panel, the left figure (B) reveals a positive link between the wealth of billionaires and their daily page views on average, indicating that billionaires who are wealthier often have more page views on average. The third plot likewise further shows a positive relationship between billionaires' wealth and the number of words in their Wikipedia articles, or in other words, that billionaires with more wealth often have longer Wikipedia articles.

For more geographical insights, data points are depicted with different shapes in the figure depending on the continent where the billionaires reside; for example, those who have citizenship in a North American country are displayed with a cross, Africa with a dot, and Asia with a triangle. As shown in the figure, the points on the very right side of each plot's axes tend to show primarily billionaires residing in North America, and we do not see many billionaires from other continents here. This pattern indicates that some billionaires from North America are outliers; they are more wealthy and more prone than billionaires from other regions of the globe to have more page views and more extensive material on their Wikipedia pages. Overall, these scatter plots show a positive link between the wealth of billionaires, the number of Wikipedia page views, and the length of Wikipedia articles, which can offer important insights into the connections between wealth, visibility, and representation in online sources.

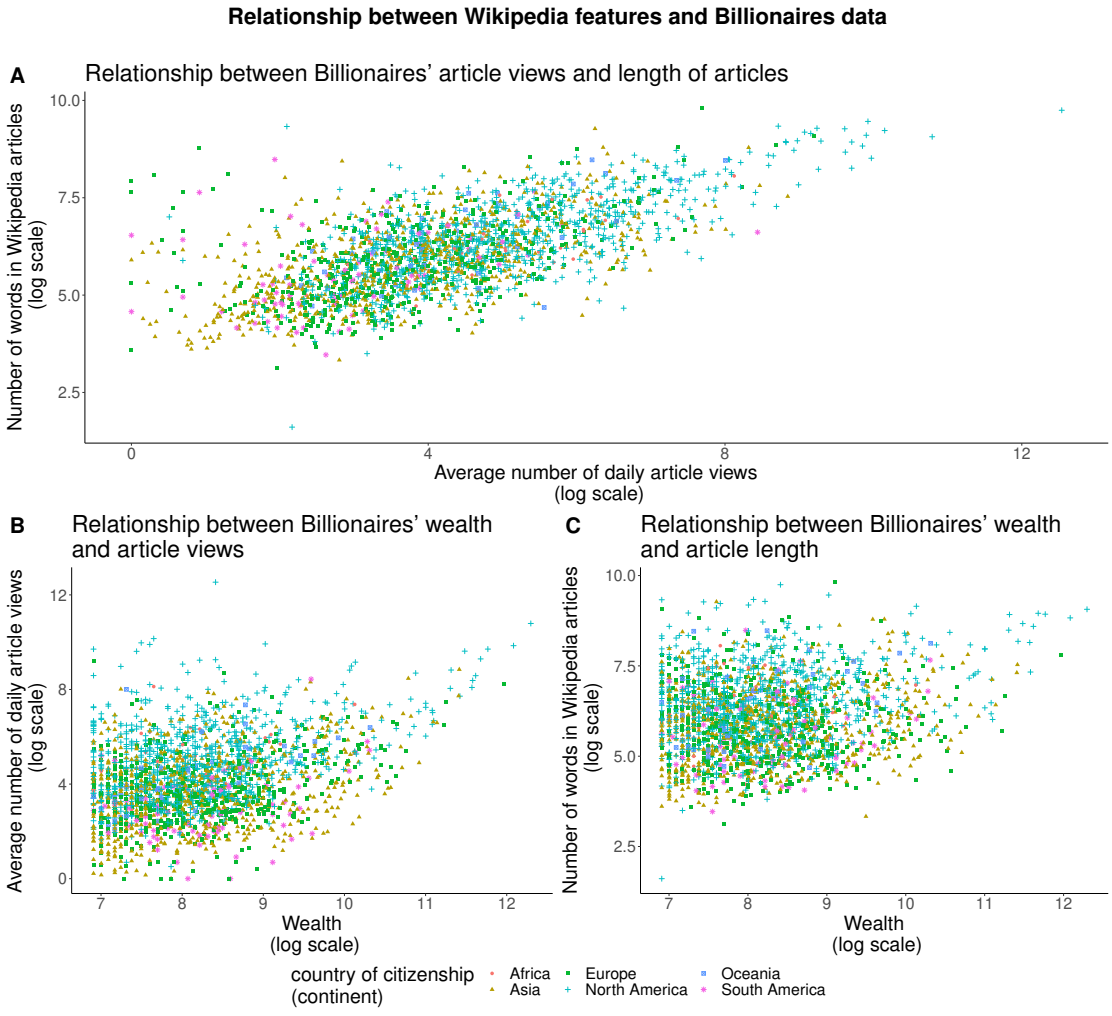

Fig. 3. Visualization of relationships between Billionaires wealth and Wikipedia features using scatter plots.

4.4. Analysing pageviews to estimate public interest in billionaires

The average daily page views of the most famous and affluent billionaires' biographies on Wikipedia are visualized in two graphs over time to see changes in public interest via Wikipedia page views (Figure 4). Panel A visualizes Elon Musk, Jeff Bezos, and Bernard Arnault as the top three wealthiest billionaires (2010-2022) and the daily page views of their Wikipedia biographies over time. The first plot demonstrates that there is relatively little

peak activity and minimal velocity in page views for Bernard Arnauld. Elon Musk and Jeff Bezos' page views, on the other hand, are comparatively larger, with considerable peaks in page views around certain dates.

Panel B in Figure 4 displays the most popular billionaires in Wikipedia that we identified beforehand with Wikipedia data on average daily page views between 2015 and the end of 2022. Over time, Donald Trump has the most page views and substantial peaks in the public interest. Although there are still some peaks in April and May 2020, Elon Musk - the most viewed among the wealthiest - has fewer page views and peaks compared to Donald Trump. Michael Jordan comes in last with the fewest page views and peaks, but with some peaks in April and May 2020.

The plots visualize fluctuations of public's interest in billionaires' over time measured by the average daily page views on Wikipedia. As shown with this application, person-related Wikipedia data may be used to measure the public interest in these billionaires, and any pattern may reveal a potential impact of their actions and/or media coverage that may cause immediate interest. For instance, peaks in Elon Musk's Wikipedia page views are particularly present at the end of April 2022, May 2021, and May 2020, which may be related to major public events such as his purchase of Twitter stocks (in April 2022), the launch of his company SpaceX (in May 2021), while May 2020 could link to some of his statements about the Covid-19 pandemic that became popular. To give another example, some of Donald Trump's peaks in Wikipedia page views were around March 2016 and November 2020, which may relate to the U.S. election, his refusal to concede and/or some of this tweets around this time, and March 2016 to his mass rally in Florida.

4.5. Network analysis with Wikidata: The family web of billionaires

Wikidata items have different properties which can be used as content for social science research. Using Wikidata's properties about family relationships such as siblings (P3373), child (P40), father (P22), mother (P25), and spouse (P26), one can scrape the whole family network of individuals included in Wikidata using the query language SPARQL. Here, the advantages of a knowledge graph become apparent. One can easily follow the different family paths and then draw and analyze the family network.

For this study, all close relatives of the billionaires are scraped (siblings, children, parents, spouses, and spouses' direct relatives). Figure 5 depicts all family webs with at least two billionaires. The filled circles represent billionaires, and the unfilled circles represent family members. The edges represent family ties between the individuals. The graph illustrates that a billionaire seldom comes alone. 590 billionaires (15% of all billionaires) are identified who have at least one other billionaire in their family. Remember that this is a lower bound because not all billionaires are included in Wikidata, and there might also be missing information on family relationships. Interestingly, some family networks are larger than nuclear families, suggesting that billionaire families are connected via marriages. This hints at social closure at the very top of the global wealth distribution. For example, Figure 6 zooms into the first family network from Figure 5. It shows a large family network, including 9 billionaires from different Indian families. The black edges represent marriage lines, and the brighter edges represent relationships between children and parents as well as siblings. The figure highlights that billionaire fortunes are connected not only via intergenerational family lines but also via marital lines.

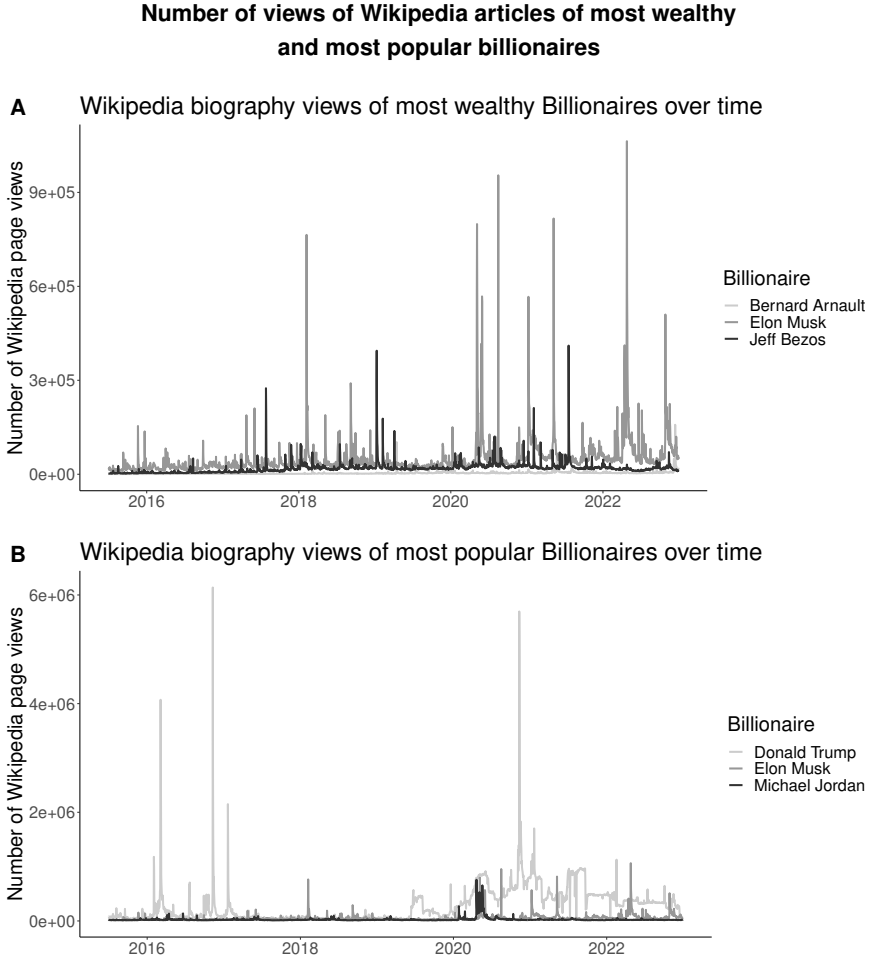

Fig. 4. Visualization of views of Wikipedia biographies of most wealthy and most popular Billionaires over time.

5. Conclusion

This study aimed at highlighting the benefits of using Wikidata and Wikipedia as research data sources for the social sciences. After summarizing the strength and weaknesses of using Wikidata and Wikipedia for social science research, a case study on global billionaires was presented to show the possibilities Wikidata and Wikipedia offer in practice.

The structured, multilingual, linked open data provided by both Wikidata and Wikipedia allow researchers to access research data or enrich their own databases. The study showed that metadata, such as the number of information provided in Wikidata items or Wikipedia articles, can be used to examine the coverage and biases of the knowledge bases. The analysis of daily article views indicates that Wikipedia can be a useful data source to study public interest in specific notable persons or subjects. Finally, the analysis of genealogical data provided in Wikidata exemplified the potential of Wikidata for social network analyses. Because Wikidata contains many different properties, it is not only possible to examine family networks but also various other networks. For example, one could identify elite institutions such as elite universities to examine how elites are connected through their educational institutions, or one could draw social networks using affiliation to political parties, cultural organizations, or companies. To conclude, Wikidata and Wikipedia can not only be used as research subjects (e.g., to examine human behavior in generating and sharing knowledge) but can also be used as content providers to do research on various topics. Furthermore, it

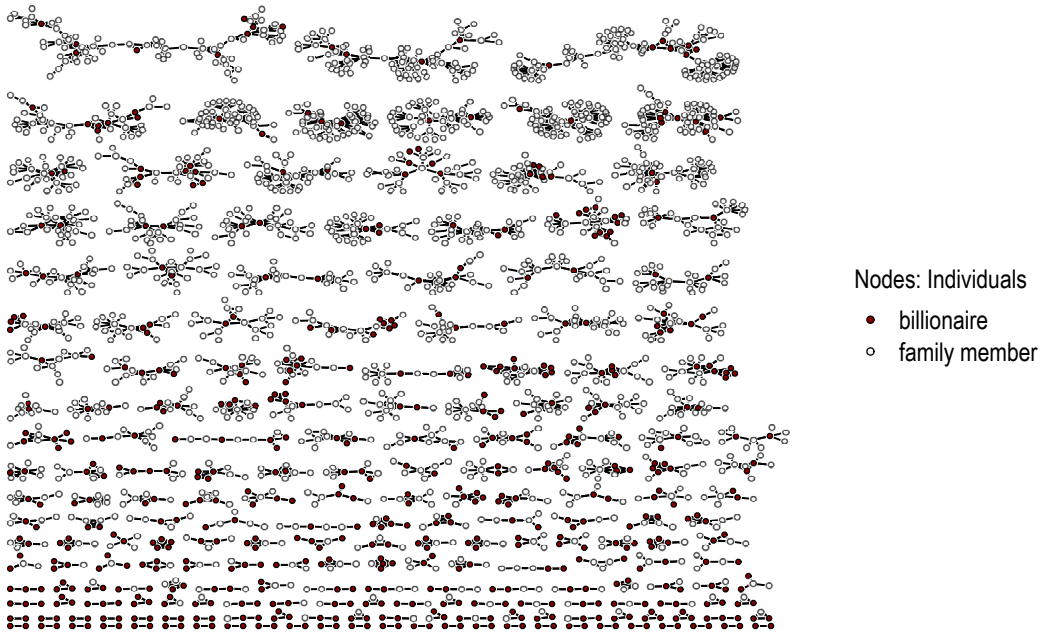

Fig. 5. Billionaire's family webs

can be used to connect one's own databases to Wikidata and to share these databases easily with other researchers, applying the good practice of open research data. Hereby, not only other researchers and the general public would benefit, but also researchers themselves may be able to advance their research by using information already included on Wikidata or other linked data.

Social scientists should also be aware of the challenges when using Wikidata and Wikipedia for research. Researchers will often be confronted with problems such as missing information and accuracy. Prior research has already highlighted the bias exhibited by knowledge graphs, but more research is needed to systematically analyze which specific groups or subjects are underrepresented in Wikidata and Wikipedia and how to best avoid biases efficiently by using different measures. If possible, researchers should complement missing data or correct incorrect or biased data. Moreover, when using Wikidata and Wikipedia for social science research, researchers need to be aware of potential biases in access, for instance, due to censorship or uneven data distribution across Wikipedia language editions. This may make comparative research challenging, and researchers need to take this into consideration. Researchers could also use a variety of sources to account for such limitations.

The case study on billionaires and Wikidata and Wikipedia provided descriptive insights for the sociological study of economic elites. The study showed that the public knowledge about billionaires is limited, with 66% of all billionaires having a Wikidata item and 57% having an English Wikipedia article. Many billionaires, thus, are not covered by Wikidata or Wikipedia, indicating the possibility of billionaires keeping their lives private. Furthermore, public knowledge tends to be biased towards North American and male billionaires. Wikipedia and Wikidata provide more information on billionaires ranking higher in the rich lists, and the number of article views is also positively related to billionaires' wealth, indicating the public's interest in the top wealth holders. Last, a brief descriptive family network analysis using genealogical data provided by Wikidata hints at social closure of the billionaires by means of intergenerational transfers and marriages.

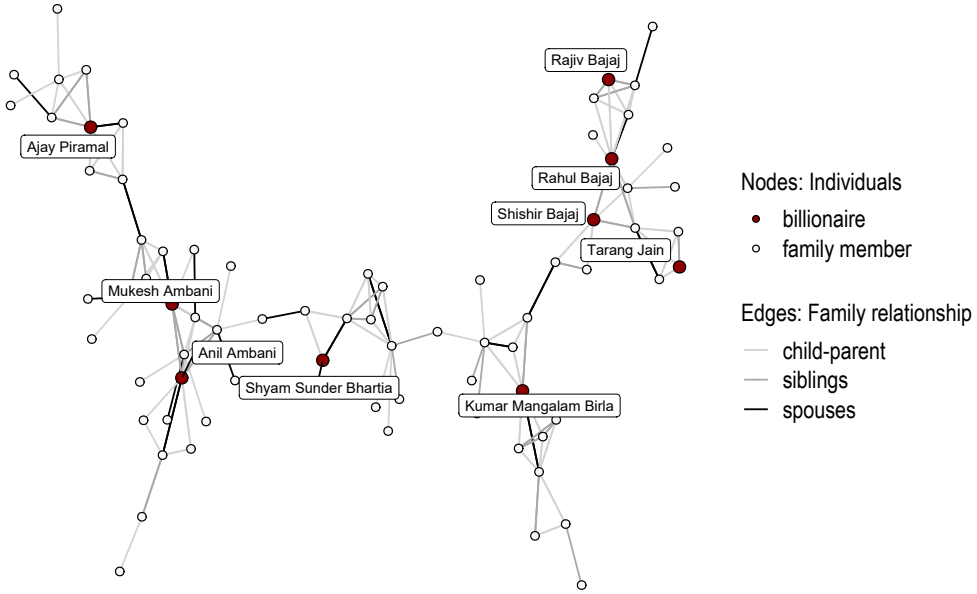

Fig. 6. Largest family web

More research is needed to uncover ways and the extent of politically motivated editing in Wikidata and Wikipedia, biographies of elites or topics, and its implications for social science research. Moreover, an in-depth analysis of the Wikipedia article content and topics in elite biographies using automated text analyses may be a fruitful direction for future research. Last, the interconnectedness of different political and economic elites and elite institutions could be analyzed using Wikidata and Wikipedia and bring valuable insights into elite research.

The opportunities for (computational) social science studies are rich and limitless, given the growing amount of data accessible in Wikidata and Wikipedia. Consequently, the future roles of Wikidata and Wikipedia in the social sciences are promising.

Acknowledgements

We thank Charlotte de Alwis for exellent research assistence and Emma Ischinsky and Ria Wilken for their efforts in linking the Forbes billionaires to Wikidata.

References

[1] M. Alam, V. De Boer, E. Daga, M. Van Erp, E. Hyvönen, and A. Meroño-Peñuela. Editorial of the Special issue on Cultural heritage and semantic web. *Semantic Web*, pages early–access, 2022. ISSN 2210-4968. Publisher: IOS Press.

[2] M. Atkinson and D. DeWitt. Does Celebrity Issue Advocacy Mobilize Issue Publics? *Political Studies*, 67(1):83–99, Feb. 2019. ISSN 0032-3217. .

[3] P. Beytía, P. Agarwal, M. Redi, and V. K. Singh. Visual gender biases in wikipedia: A systematic evaluation across the ten most spoken languages. volume 16, pages 43–54, 2022. ISBN 2334-0770.

[4] B. Castanho Silva and S.-O. Proksch. Politicians unleashed? Political communication on Twitter and in parliament in Western Europe. *Political Science Research and Methods*, 10(4):776–792, July 2022. ISSN 2049-8470. . URL https://www.cambridge.org/core/article/politicians-unleashed-political-communication-on-twitter-and-in-parliament-in-western-europe/AF03501C445E752094C9B380C4E4913E. Edition: 2021/07/15 Publisher: Cambridge University Press.

[5] J. Clark, R. Faris, and R. Heacock Jones. Analyzing accessibility of Wikipedia projects around the world. *Berkman Klein Center Research Publication*, (2017-4), 2017.

[6] D. Cohen, N. Baumann, and S. Munzert. Studying Politics on and with Wikipedia, Methods Bites - Blog of the MZES Social Science Data Lab, 2019. URL https://www.mzes.uni-mannheim.de/socialsciencedatalab/article/studying-politics-wikipedia/.

[7] M. Elshendy, A. F. Colladon, E. Battistoni, and P. A. Gloor. Using four different online media sources to forecast the crude oil price. *Journal of Information Science*, 44(3):408–421, June 2018. ISSN 0165-5515, 1741-6485. . URL http://journals.sagepub.com/doi/10.1177/0165551517698298.

[8] M. Farda-Sarbas and C. Müller-Birn. Wikidata from a Research Perspective – A Systematic Mapping Study of Wikidata. *arXiv*, 2019. URL

[9] J. Geiß and M. Gertz. With a little help from my neighbors: person name linking using the Wikipedia social network. pages 985–990, 2016.

[10] R. Gieck, H.-M. Kinnunen, Y. Li, M. Moghaddam, F. Pradel, P. A. Gloor, M. Paasivaara, and M. P. Zylka. Cultural differences in the understanding of history on Wikipedia. pages 3–12. Springer, 2016. ISBN 3-319-42696-6.

[11] P. Gloor, P. De Boer, W. Lo, S. Wagner, K. Nemoto, and H. Fuehres. Cultural anthropology through the lens of Wikipedia-A comparison of historical leadership networks in the English, Chinese, Japanese and German Wikipedia. *arXiv preprint arXiv:1502.05256*, 2015.

[12] S. Gobel and S. Munzert. The Comparative Legislators Database. *British Journal of Political Science*, 52(3):1398–1408, July 2022. ISSN 0007-1234. . Number: 3.

[13] S. Göbel and S. Munzert. Political Advertising on the Wikipedia Marketplace of Information. *Social Science Computer Review*, 36(2): 157–175, Apr. 2018. ISSN 0894-4393, 1552-8286. .

[14] O. Keyes and B. Tilbert. WikipediR: A MediaWiki API Wrapper. 2017. Publisher: R package version.

[15] S. R. Khan. The Sociology of Elites. *Annual Review of Sociology*, 38(1):361–377, 2012. ISSN 0360-0572.

[16] M. Klein, H. Gupta, V. Rai, P. Konieczny, H. Zhu, and ACM. Monitoring the Gender Gap with Wikidata Human Gender Indicators. In *University of Minnesota System*, Berlin, Germany, 2016. ISBN 978-1-4503-4451-7. .

[17] N. T. Korfiatis, M. Poulos, and G. Bokos. Evaluating authoritative sources using social networks: an insight from Wikipedia. *Online Information Review*, 30(3):252–262, 2006. ISSN 1468-4527. Publisher: Emerald Group Publishing Limited.

[18] P. Korom, M. Lutter, and J. Beckert. The enduring importance of family wealth: Evidence from the Forbes 400, 1982 to 2013. *Social Science Research*, 65:75–95, 2017.

[19] P. Laufer, C. Wagner, F. Flöck, and M. Strohmaier. Mining cross-cultural relations from Wikipedia: a study of 31 European food cultures. pages 1–10, 2015.

[20] M. Lutter, K. L. A. Roex, and D. Tisch. Anomie or imitation? The Werther effect of celebrity suicides on suicide rates in 34 OECD countries, 1960-2014. *Social Science & Medicine*, 246:112755, Feb. 2020. ISSN 0277-9536. . URL

[21] N. Nazeri and C. Anderson. Citation filtered: Iran's censorship of Wikipedia. 2013.

[22] F. Pradel. Biased Representation of Politicians in Google and Wikipedia Search? The Joint Effect of Party Identity, Gender Identity and Elections. *Political Communication*, 38(4):447–478, July 2021. ISSN 1058-4609. . Number: 4.

[23] E. Seidlmayer, J. Voß, T. Melnychuk, L. Galke, K. Tochtermann, C. Schultz, and K. U. Förstner. ORCID for Wikidata. Data enrichment for scientometric applications. CEUR Workshop Proceedings, 2020.

[24] T. Shafee, O. Keyes, S. Signorelli, A. Lum, C. Graul, and M. Popov. WikidataR: API client library for'Wikidata'. 2022. URL https://github.com/TS404/WikidataR. Publisher: R package version.

[25] E. Shor, A. Van De Rijt, and B. Fotouhi. A large-scale test of gender bias in the media. *Sociological Science*, 6:526–550, 2019. ISSN 2330-6696.

[26] M. Toft and V. Jarness. Upper-class romance: homogamy at the apex of the class structure. *European Societies*, 23(1):71–97, Jan. 2021. ISSN 1461-6696. . URL https://doi.org/10.1080/14616696.2020.1823009https://www.tandfonline.com/doi/pdf/10.1080/14616696.2020.1823009?needAccess=true.

[27] C. Wagner, D. Garcia, M. Jadidi, and M. Strohmaier. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, pages 454–463, 2015. URL https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/viewFile/10585/10528.

[28] C. Wagner, E. Graells-Garrido, D. Garcia, and F. Menczer. Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Science*, 5:1–24, 2016. Publisher: Springer.

[29] Wikidata. Wikidata: Introduction. 2023. URL https://www.wikidata.org/wiki/Wikidata:Introduction.

[30] Wikipedia. Censorship of Wikipedia, Feb. 2023. URL https://en.wikipedia.org/w/index.php?title=Censorship_of_Wikipedia&oldid=1138760104. Page Version ID: 1138760104.

[31] Wikipedia. Wikipedia:Systemic bias, Feb. 2023. URL https://en.wikipedia.org/w/index.php?title=Wikipedia:Systemic_bias&oldid=1138230255. Page Version ID: 1138230255.

[32] O. F. Yalcin. Measuring and Modeling the Dynamics of Elite Political Networks. 2021. Publisher: Pennsylvania State University.

[33] E. Yang and M. E. Roberts. Censorship of online encyclopedias: Implications for NLP models. pages 537–548, 2021.

[34] T. Yasseri, A. Spoerri, M. Graham, and J. Kertész. The most controversial topics in Wikipedia. *Global Wikipedia: International and cross-cultural issues in online collaboration*, 25:25–48, 2014.

[35] R. Zakharenko. Dead men tell no tales: the role of cultural transmission in demographic change. *Journal of Demographic Economics*, 87(4):511–536, 2021. ISSN 2054-0892. . URL https://www.cambridge.org/core/article/dead-men-tell-no-tales-the-role-of-cultural-transmission-in-demographic-change/957004BAE863B1F14F980331B2DB7BA8. Number: 4 Edition: 2021/05/18 Publisher: Cambridge University Press.

[36] F. Zhao. A systematic review of Wikidata in Digital Humanities projects. *Digital Scholarship in the Humanities*, pages 1–23, Dec. 2022. ISSN 2055-7671, 2055-768X. . URL https://academic.oup.com/dsh/advance-article/doi/10.1093/llc/fqac083/6964525.