Using the Web of Data to Study Gender Differences in Online Knowledge Sources: the Case of the European Parliament

Laura Hollink
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
l.hollink@cwi.nl

Astrid van Aggelen
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
aggelen@cwi.nl

Jacco van Ossenbruggen
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
Jacco.van.Ossenbruggen@cwi.nl

ABSTRACT

Gender inequalities are known to exist in Wikipedia. However, objective measures of inequality are hard to obtain, especially when comparing across languages. We study gender differences in the various Wikipedia language editions with respect to coverage of the Members of the European Parliament. This topic allows a relatively fair comparison of coverage between the (European) language editions of Wikipedia. Moreover, the availability of open data about this group allows us to relate measures of Wikipedia coverage to objective measures of their notable actions in the offline world. In addition, we measure gender differences in the content of Wikidata entries, which aggregate content from across Wikipedia language editions.

CCS CONCEPTS

• **Human-centered computing** → **Wikis**; • **Social and professional topics** → **Gender**; • **Computing methodologies** → *Knowledge representation and reasoning*; • **Information systems** → *World Wide Web*;

KEYWORDS

Web of Data, Wikipedia, Wikidata, Gender Inequality, European Parliament

ACM Reference Format:
Laura Hollink, Astrid van Aggelen, and Jacco van Ossenbruggen. 2018. Using the Web of Data to Study Gender Differences in Online Knowledge Sources: the Case of the European Parliament. In *WebSci '18: 10th ACM Conference on Web Science, May 27–30, 2018, Amsterdam, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3201064.3201108

1 INTRODUCTION

We study gender differences in Wikipedia language editions with respect to how members of the European Parliament are covered. Gender differences are known to exist in Wikipedia. Studies have uncovered gender inequalities with respect to the coverage of people [6, 10], the textual and structural content of the articles [9, 10], and the quality [4] of the articles. Given the scale at which Wikipedia is accessed[1] and used as a source of knowledge, it is important to reliably assess to what extent and in what way the content of this collaboratively edited online encyclopedia might be biased.

Objective measures of inequality are hard to obtain. The difficulty stems from the fact that we don't know who (or what) has not been covered. Several studies have used reference corpora of notable people to measure what is missing from Wikipedia [6, 9]. The downside of this approach is that the reference corpus itself might be biased. In addition, the two groups of people that are to be compared (in this case men and women) might differ in ways that make them hard to compare. For example, people of different genders may have had different occupations or roles, and the quality of their historic records may be different. Wagner et al. [10] use two metrics to estimate how "notable" a person in a reference corpus is. They find that coverage bias with respect to gender is small among highly notable people but large among less-notable people. In other words, women need to be very notable to be included in Wikipedia, while little-notable men may be included relatively easily.

Cross-language comparisons of Wikipedia editions are even more complicated, since they involve a comparison of different (but overlapping) groups of people: a person who is relevant to one language community might not meet the criteria for inclusion for another community. As a result, most studies focus on one or a few languages.

In this study, we focus on Wikipedia coverage of a relatively small and homogeneous group of people that is of European-wide importance: the 3662 (past and current) Members of the European Parliament (MEPs). This group is relevant for all member states of the European Union, and thus allows a relatively fair comparison of coverage between the (European) language editions of Wikipedia. Moreover, the availability of open data about this group of people provides us with objective data of their notable actions in the European Parliament. Thus, we can relate the observed differences in how men and women are presented on Wikipedia to their characteristics and actions in the offline world. We use various sources of open data, collected and published as Linked Open Data in the Talk of Europe project[2]. We use Wikidata to connect all Wikipedia editions to each other and to the open data about the European Parliament, and to get reliable data about the gender of MEPs.

2 RELATED WORK AND MOTIVATION

Wagner et al. [9] distinguished four types of gender-bias on Wikipedia: visibility, structural, lexical and coverage bias. They analyzed bias in six Wikipedia language editions: English, German, French, Spanish, Italian and Russian. They found that men and women are equally

[1] See e.g. https://www.alexa.com/siteinfo/wikipedia.org for access rates.
[2] http://talkofeurope.eu/

WebSci '18, May 27–30, 2018, Amsterdam, Netherlands

ACM ISBN 978-1-4503-5563-6/18/05…$15.00
https://doi.org/10.1145/3201064.3201108

visible; that a structural bias exists in that men are on average more central (in all language editions except Spanish); and that a lexical bias exists in the content of biographic articles, where woman's biographies contain more words related to family and relationships. The latter finding is a confirmation of Bamman and Smith [1], who also analyzed Wikipedia article texts and found that articles about women contain more information about marriage and divorce. Graells-Garrido et al. [3] studied differences in content between articles about men and women as well. Next to text analysis, they included structural properties from (English) DBpedia[3] entries, and found that the representation of women varies greatly between DBpedia categories with, for instance, an over-representation of women in the Artist and Model categories but an over-representation of men in the Athlete and Politician categories. Halfaker [4] studied article quality on Wikipedia. He shows that article quality for women scientists was initially below average, but that since the start of targeted efforts to increase the coverage of women on Wikipedia the opposite effect is now true: women scientists' articles have an above-average quality.

We complement the above work in a number of ways. First, we include 23 Wikipedia language editions where previous studies have focused on one or a few languages. Second, we include Wikidata [8] in our analysis. This online knowledge base contains structured knowledge that is, in many cases, aggregated from various Wikipedia language editions. Third, our focus on one specific category, namely members of the European Parliament, allows us to uncover biases which may remain invisible in general analyses over all categories, as [3] remarks. The choice for a European Union-wide topic allows a (relatively) fair comparison of coverage across (European) language editions. The fact that we have a complete reference corpus, including data on the notable actions of the men and women involved, solves many of the difficulties that other studies face with respect to measuring coverage bias.

On the other hand, the present study is several orders of magnitude smaller than the studies described above, and is limited to an investigation of coverage and content, where content is studied strictly in terms of Wikidata properties (in contrast to e.g. a textual analysis of Wikipedia article content).

3 DATA AND METHODS

Data about the European Parliament was taken from the Talk of Europe project [7]. In Talk of Europe, various Web sources about the European Parliament were combined and translated into the Web format RDF. This data includes, for example, information about the speeches that MEPs perform in the plenary debates of the parliament, crawled from the website of the EP[4], and information about professional affiliations of the members of the EP [5], including committee membership and the roles that people have (chair, vice-chair, substitute, member of the bureau, etc.). The Talk of Europe data includes links to Wikidata entries for MEPs, which were generated based on unique identification numbers used on both the EU website and Wikidata.

From Wikidata we retrieved all entries of Members of the European Parliament. This data consists of the property-value pairs associated to each MEP, including the Gender property, as well as links to the Wikipedia language editions that cover them. We limit the analysis to language editions that correspond to 23 languages spoken at the European Parliament[5]. Wikidata includes several genders, but only the values "male" and "female" have enough occurrences among the MEPs to allow a valid comparison. We include one "transgender female" MEP in our analysis as "female".

We examine differences between male and female MEPs with respect to (1) the number of Wikipedia language editions that cover them, (2) which Wikipedia language editions cover them, and (3) the number of property-value pairs in their Wikidata entries. In addition (4), we examine the properties that are typical for either one of the two genders. Next to raw frequency counts, we use Pointwise Mutual Information (PMI) [2] as a measure for how typical a property is for a gender. The PMI value of a property Y for a gender X is defined as:

$$PMI(X, Y) = \log \frac{P(X, Y)}{P(X)P(Y)}$$

where the value of $P(X)$ is the proportion of gender X in the data and $P(X, Y)$ is the proportion MEPs who's Wikidata entry includes property Y.

We relate the observed gender differences in Wikipedia and Wikidata to characteristics and actions of the MEPs in the European Parliament. Specifically, we look at (a) the amount of male and female representatives of each country in the Parliament, (b) the number of speeches of each MEP in the Parliament, and (c) the number of chair-positions that the MEPs held within the European Parliament. The latter two can be seen as external measures of notability. We use them to test whether we can replicate the findings by [10] that women need a higher notability to be included in Wikipedia than men. Note that a one-to-one mapping between countries of the European Union and Wikipedia language editions is not always possible. Some languages are spoken in multiple countries (e.g. English and German) and some countries have multiple official languages (e.g. Luxembourg and Austria).

Throughout the paper we have used the nonparametric Wilcoxon rank sum test to compare men and women, and Spearman's rank correlation coefficient (ρ), which is robust to outliers, for correlations between variables.

4 ANALYSIS AND RESULTS

4.1 Number of Wikipedia Language editions

Female members of the European Parliament are covered by slightly more Wikipedia language editions than male members ($p < 0.01$, Figure 1). The 969 female MEPs have a median of 5 and mean of 5.9 Wikipedias; the 2693 male MEPs have a median of 4 and mean of 6.4. This is in line with earlier studies (e.g. [10]), who found that women on Wikipedia - at least those born since 1900 - had a higher Wikipedia language edition count then men.

There is a relation between Wikipedia presence and an MEP's effort as recorded in the European Parliament. We found a positive correlation between the amount of speeches that a person held in parliament and the number of Wikipedia language editions that

[3] http://dbpedia.org/

[4] http://www.europarl.europa.eu, with speech data available since 1999

[5] See http://www.europarl.europa.eu/aboutparliament/en/20150201PVL00013/Multilingualism. We don't have speech data in Gaelic.

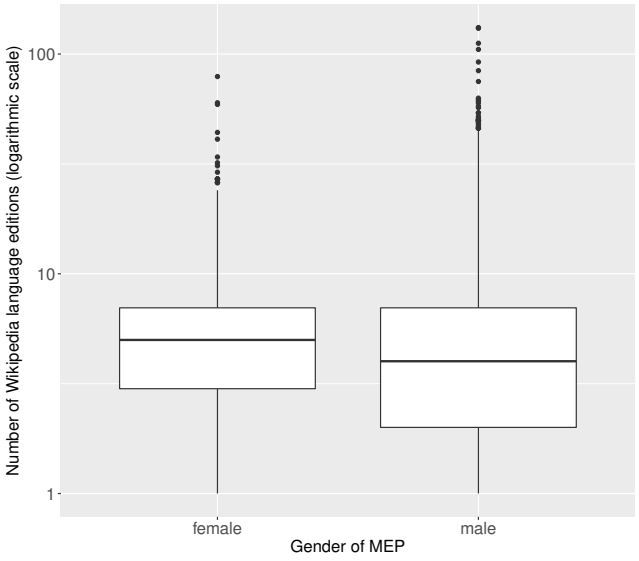

Figure 1: Boxplot of the number of Wikipedia language editions that cover female and male MEPs.

cover them (ρ =0.54, p < 0.01). When we split the MEPs into those that spoke a lot (>= 100 times) and those that spoke less (<100 times), we find that those who spoke a lot have more Wikipedia presence - a median of 6 Wikipedias vs. 4 for those who spoke less. The same effect is present for MEPs who have held chair-positions within the EP vs. those that were never chairs (5 and 4 Wikipedias, respectively). Table 1 lists the median number of Wikipedias for each group per gender.

The observed effect that women are covered by more Wikipedias is only present in the groups with low effort recorded: women have more Wikipedias than men in the groups that speak less or were never chairs (p < 0.01 for both), but there is no significant difference between the two genders in the groups of MEPs that speak a lot or held chair positions (p = 0.4 and 0.28, respectively),

These results are different from what was found in an earlier study over all Wikipedia biographies by Wagner et al. [10]. They use the number of Wikipedia language editions as an (internal) measure of notability and observe a larger negative bias against women in the group of less notable people. They conclude that women need a higher notability to be included in Wikipedia than men. If we treat the number of speeches and chair-positions as (external) measures of notability, we would expect a negative bias against women in the *spoke-less* group and in the *nonchairs* group. This is not the case in our data and, hence, we cannot confirm that female MEPs need a higher notability to be included in a Wikipedia edition than male MEPs. Further research on other measures of notability is needed to determine the solidity of this conclusion.

4.2 Variation in Wikipedia Language editions

We found large differences between Wikipedia language editions with respect to overall coverage of MEPs and gender balance. Some editions over-represent women, while others over-represent men. Figure 2 shows the percentage of female and male MEPs that are

Table 1: Median Wikipedia language editions of MEPs.

MEPs	All	female	male
All	4	5	4
Spoke-a-lot	6	6	6
Spoke-less	4	4	3
Chairs	5	6	5
Nonchairs	4	5	4

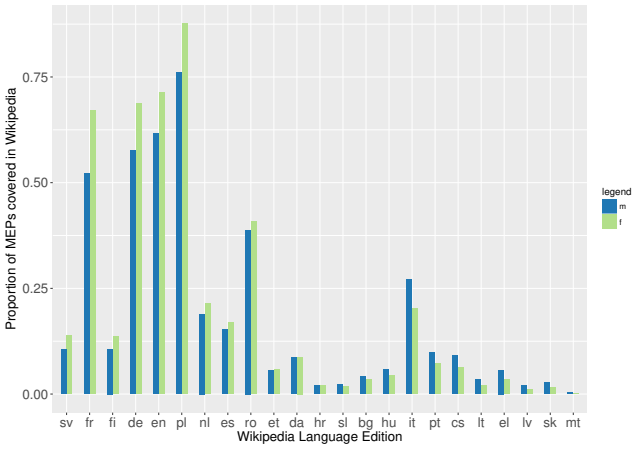

Figure 2: The proportion of female and male MEPs that are covered in 23 Wikipedia language editions.

covered in each edition. None of the editions cover 100% of the MEPs. Ten language editions include a higher percentage of women than men. Thirteen editions cover a lower percentage of female MEPs than male MEPs.

There is a strong correlation between completeness of the edition and the size of the imbalance[6] (ρ = 0.75, p < 0.01). This may suggest that biographies of males are prioritized by the Wikipedia editor communities, but further study on Wikipedia edit histories is needed to confirm this hypothesis.

The observed differences between Wikipedia language editions are in line with the gender-balances among representatives of the member states in the EP. Figure 3 plots the number of MEPs from 1999 to 2017 for the countries whose Wikipedia language edition[7] most strongly over-represents women (left) or men (right). While the numbers vary per country and over time, the data suggest that the countries with Wikipedias that over-represent women have a higher percentage of female MEPs than the countries whose Wikipedia over-represents men. For completeness, table 2 lists the

[6] completeness is operationalized as the percentage of MEPs covered, and the size of the imbalance as the percentage of female MEPs covered divided by the percentage of male MEPs covered

[7] As mentioned in section 3, a one-to-one mapping between countries and Wikipedia language editions is not always possible. For the ten countries in Figure 3, a mapping was created manually. We have left the UK out of this figure because we assume that the English Wikipedia is edited by a wider community.

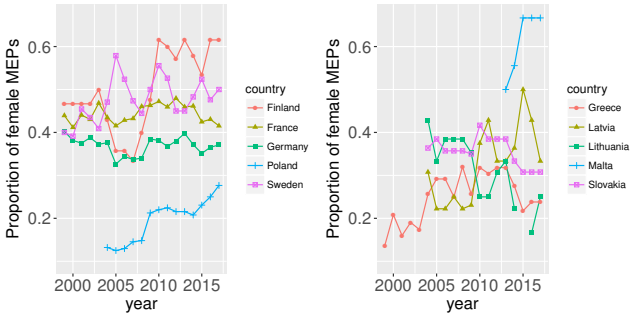

Figure 3: Proportion of female MEPs for countries whose Wikipedia over-represents women (left) or men (right).

Table 2: Nr. of female and male MEPs per country in 2017.

Country	f	m	Country	f	m
Malta	4	2	Belgium	7	12
Finland	8	5	Latvia	2	4
Ireland	5	4	Slovakia	4	9
Croatia	6	5	Portugal	6	14
Austria	9	9	Romania	9	22
Estonia	3	3	Poland	13	34
Sweden	10	10	Czech Republic	5	15
Spain	24	29	Denmark	3	9
France	27	38	Lithuania	2	6
Netherlands	11	16	Greece	5	16
United Kingdom	26	39	Bulgaria	4	13
Slovenia	3	5	Hungary	4	16
Germany	33	56	Luxembourg	1	4
Italy	24	41	Cyprus	1	5

number of female and male MEPs for all countries in 2017, ordered by decreasing proportion of female representatives[8].

4.3 Properties in Wikidata entries

We found no difference between the *number* of property-value pairs in Wikidata entries of male and female MEPs (p = 0.09, Figure 4). Both have a median of 20 properties (Table 3). Small gender-differences are visible within specific groups of MEPs but these are not significant (only the small difference of 19 vs. 20 properties observed within the spoke-less group is significant with p = 0.03).

We also inspected differences in the *content* of Wikidata entries of male and female MEPs. Raw counts of property occurrence do not reveal these differences. For example, the top 10 most frequent properties for female MEPs is identical to the top 10 for males except for slight differences in ranking. When inspecting properties with a high PMI value for male or female MEPs - in other words, those properties that are typical for either of the two genders - differences become visible. To characterize these differences, we manually identified properties that relate to relationships or family

[8]The representatives change periodically with EP elections as well as incidentally throughout parliamentary years. We have included the 702 MEPs who spoke in parliament in the year 2017.

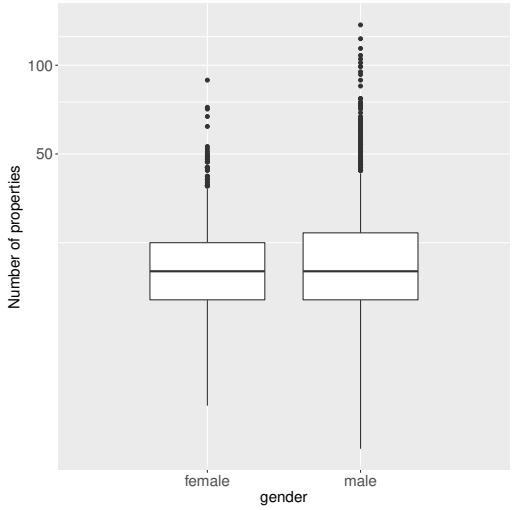

Figure 4: Number of properties per Wikidata entry for MEPs.

Table 3: Median Wikidata properties in entries of MEPs.

MEPs	All	female	male
All	20	20	20
Spoke-a-lot	21	21	21
Spoke-less	20	19	20
Chairs	24	22	24
Nonchairs	20	19	20

(cf. [10]). We only tagged the properties that occur at least once in a male entry and a female entry, aiming to exclude properties that are not applicable to one of the genders. This was the case for 224 of the 396 distinct properties that appear in our dataset.

Family/relationship-properties with a positive PMI value for female MEPs are (in order of PMI) "spouse", "father" and "number of children." The family/relationship-properties typical for men are "partner", "child", "sibling", "family", "relative", "sexual orientation" and "mother." Some of these properties have a frequency of occurrence that is too low to draw valid conclusions. For example, "sexual orientation" is recorded for only four MEPs. Table 4 lists the top 20 properties with highest PMI values for male or female MEPs that occur at least 30 times. Properties relating to classifications schemes in use in a particular country, which may be unfamiliar to the reader, are identified as such.

Based on these data, we see no evidence that Wikidata entries of female MEPs contain more relationship- or family-related content than entries of male MEPs. Rather, the differences in content seem to be related to nationality. Typical properties for men are used in countries that over-represent men in the EP (e.g. Czech Republic, Poland, Latvia) and/or in countries whose Wikipedia over-represents men (Latvia, Czech Republic, Italy). The properties most typical for women include identifiers from classifications schemes in Finland, Sweden, the U.K., Austria, France and Germany

Table 4: Properties with highest PMI values for men and women that occur in at least 30 Wikidata entries of MEPs.

Property	Freq. m	Freq. f	National Class. scheme?
Highest PMI value for men:			
topic's main category	30	2	
described by source	39	3	
NLA (Australia) ID	37	3	Australia
signature	48	4	
National Library of Israel ID	32	3	Israel
SHARE Catalogue author ID	41	4	Italy
NDL Auth ID	51	5	Japan
Gran Enciclop. Catalana ID	81	8	Catalonia
Encyclop. Univers. Online ID	59	6	France
BAV ID	29	3	Italy (Vatican)
Perlentaucher ID	29	3	Germany
PM20 folder ID	48	5	Germany
Encyclop. Britannica Online ID	76	8	
NLP ID	45	5	Poland
child	154	18	
CANTIC-ID	84	10	Catalonia
NNDB people ID	41	5	
NKCR AUT ID	76	10	Czech rep.
LNB ID	30	4	Latvia
place of death	354	48	
Highest PMI value for women:			
Finnish MP ID	23	27	Finland
birth name	51	51	
Riksdagen person-id	36	31	Sweden
residence	36	30	
Google Knowledge Graph ID	22	16	
parliament.uk ID	25	18	U.K.
video	62	43	
spouse	90	55	
Twitter username	240	142	
Austrian Parliament ID	46	27	Austria
Facebook profile ID	140	73	
SELIBR Id	51	25	Sweden
Babelio author ID	25	12	France
biogr. at the Bundestag of DE	23	11	Germany
Who's Who in France biogr. ID	142	67	France
official website	387	178	
Commons category	1011	451	
image	1360	600	
participant of	185	81	
IPA transcription	30	13	

- countries with more women in the EP than average, and whose Wikipedias over-represents women MEPs.

Another possible cause of the observed differences is age. The fact that "child" is typical for entries of men could be due to the fact that in the past the proportion of male MEPs was larger than in recent years. Therefore, the average age of the male MEPs is higher, making it more likely that their children are notable enough to appear on Wikidata. This also explains why "place of death" is typical for men. The relatively young group of female MEPs is more likely to have records of their Twitter and Facebook IDs.

5 DISCUSSION AND CONCLUSION

We found a very small gender-difference in the number of Wikipedia language editions that cover an MEP, with women covered by slightly more editions. The variation among language editions is large, with some editions over-representing women and others over-representing men. The inequality in a Wikipedia edition seems to correspond to the gender (im)balance among the representatives of the nations in the EP. It may suggest a larger gender inequality in those countries, but further studies are necessary.

Male and female MEPs have a similar number of property-value pairs in the Wikidata entries, but we found differences in the content of the entries. We could not reproduce findings of previous studies regarding the content of male and female (Wikipedia) entries, namely that those of women over-emphasize family and relationship topics. Rather, in our data, the differences seem to relate to differences in the real world, namely gender imbalance among the EP representatives of the various member states, and birth year of the MEPs.

Several things could play a role here. First, it is possible that a subtle bias is noticeable in natural language text of Wikipedia articles but not in the structured data that we analyzed. Second, our study is limited to one profession and includes only fairly notable people. Previous studies showed large differences between professions and between notable and less-notable people, which may explain why we could not reproduce their findings. Finally, the fact that Wikidata aggregates information from many Wikipedias might help diversity and decrease inequality of the content. This would be a very positive effect of the efforts of Wikidata editors and developers.

ACKNOWLEDGMENTS

We thank Jan Wielemaker for providing help with data analysis on SWISH DataLab. This research was partially supported by the VRE4EIC project, funded from H2020 grant No 676247.

REFERENCES

[1] David Bamman and Noah A. Smith. 2014. Unsupervised Discovery of Biographical Structure from Text. *TACL* 2 (2014), 363–376. https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/371

[2] Kenneth Ward Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16, 1 (1990), 22–29.

[3] Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. 2015. First women, second sex: gender bias in Wikipedia. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM, 165–174.

[4] Aaron Halfaker. 2017. Interpolating Quality Dynamics in Wikipedia and Demonstrating the Keilana Effect. In *Proceedings of the 13th International Symposium on Open Collaboration*. ACM, 19.

[5] Bjørn Høyland, Indraneel Sircar, and Simon Hix. 2009. Forum section: an automated database of the european parliament. *European Union Politics* 10, 1 (2009), 143–152.

[6] Joseph Reagle and Lauren Rhue. 2011. Gender bias in Wikipedia and Britannica. *International Journal of Communication* 5 (2011), 21.

[7] Astrid van Aggelen, Laura Hollink, Max Kemman, Martijn Kleppe, and Henri Beunders. 2017. The debates of the European Parliament as Linked Open Data. *Semantic Web* 8, 2 (2017), 271–281. DOI: http://dx.doi.org/10.3233/SW-160227

[8] Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85. DOI: http://dx.doi.org/10.1145/2629489

[9] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia.. In *ICWSM*. 454–463.

[10] Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. 2016. Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Science* 5, 1 (2016), 5.