Quality vs. Quantity:

A Metadata Analysis of Gender Representation on Wikipedia

Brandon D. Peri, Wilbert Osmond, Nienke E. Legemaate, Florian M. Kleinau

Quality vs. Quantity:
A Metadata Analysis of Gender Representation on Wikipedia

Introduction

Wikipedia is an open-access, multilingual, collaborative online encyclopedia curated and maintained by a community of volunteer editors. Since its inception in 2001, Wikipedia has become an increasingly prevalent source of information used by both academics and laymen alike (Dutta et al., 2008; Singer et al., 2017). However, as Wikipedia has become increasingly important as an educational tool, so too has it become increasingly critical for its database to feature accurate and equitable representation of the social groups represented on it.

Thus far, current research suggests that Wikipedia has failed to do so. Previous studies have found that women are underrepresented on Wikipedia, with female entries making up 15.53% of biographies in 2014, and (even after active effort to increase this number) 18.71% in 2021 (Women-in-Red, n.d.). In addition, differences in the meta-data, language, and network structures of male and female biographies have been found, with women being linked more to male pages than vice-versa, and female pages containing more references to familial issues and romantic relationships than male ones (Graells-Garrido et al., 2015; Wagner et al., 2015).

In this paper, we aim to further investigate the extent to which gender bias is present in Wikipedia's database by providing a novel quantitative analysis of differences in its representation of men and women. In line with previous studies, we expect to find that women are not only underrepresented (i.e. the ratio of male and female articles does not correspond to the male/female ratio in the human population), but *misrepresented* on Wikipedia, with female articles reflecting archaic gender stereotypes in which their relationships to others are valued more highly than their own achievements. We therefore hypothesize that male biographies will more frequently contain information on their education and occupation, whereas female entries will more frequently feature information on family and relationships.

Methods

Data

Our analyses employed two publicly available databases:

1. **DBPedia (Lehmann et al., 2015)**. Our primary data-set is the DBPedia project which consisted of sets of algorithmically extracted metadata information about the contents of Wikipedia pages. For our analysis, we used DBPedia's 'The People of Wikipedia' data-set, which contained an array of information for over 1.5 million biographical Wikipedia articles (Lehmann et al., 2015). The meta-data included in this data-set was collected mainly by detecting info boxes in biographies and assigning each piece of information from these boxes to a respective subclass. The original data set was accessed on January 18th, 2021.

2. **Inferred gender for Wikipedia biographies (Bamman & Smith, 2014)**. In order to supplement the rather sparse gender data present in the DBPedia data-set, we employed a secondary data-set which contained the inferred gender for 862,171 people on Wikipedia (Bamman & Smith, 2014). The inferred gender was assigned based on the number of grammatically gendered words present in the biographies. As the genders considered in this set were strictly male or female, these were the sole categories of gender considered in our analyses. This data set was also accessed on January 18th, 2021.

Pre-processing

We pre-processed the DBPedia (Lehmann et al., 2015) data set by removing all elements for which 'fictional character' was defined as a meta-data attribute, and only included entries for which 'birth year' or 'birth date' was reported. Doing so reduced the number of entries considered to 975,235. We then merged this set with the "Inferred gender" (Bamman & Smith, 2014) data set (removing all DBPedia (Lehmann et al., 2015) entries which were not included in the Bamman and Smith (2014) set), which further reduced our final data set to include 631,258 elements.

Analysis

To assess systematic differences in gender representation, we computed the relative prevalence of various meta-data attributes in male and female entries. By simply observing the presence or absence of certain attributes (e.g. whether or not children were reported), we argue that robust insights can be drawn about the contents of their respective articles, without needing to perform more complex linguistic analysis. We subdivided the meta-data attributes considered in our analysis into one of several categories:

1. Education (almaMater_label, education_label)
2. Occupation (occupation_label, profession_label)
3. Net Worth (networth_label)
4. Known For (knownFor_label)
5. Family (relation_label, relative_label, spouse_label, child_label, parent_label)

Out of the metadata attributes defined by the DBPedia (Lehmann et al., 2015) data set, we argue that these are most indicative of conformity to archaic gender stereotypes - where men are traditionally seen as educated, autonomous, and professional, and where women are considered primarily in relation to their children, fathers, and spouses (Wagner et al., 2015). By comparing male and female Wikipedia entries on these categories, we aimed to reveal the extent to which the (notable) men and women featured in their database reflect these traditional views on gender.

As initial, exploratory analysis revealed that almaMater_label and education_label were often used interchangeably throughout our data-set, we chose to simply sum the number of occurrences of these attributes. Likewise, occupation_label and profession_label were similarly aggregated.

In addition to the primary analysis computed over our aggregated meta-data categories, we also conducted a separate analysis of the relative prevalence for each of the metadata attributes included in our 'Family' category. Unlike for the attributes

included in occupation and education, are not interchangeable, and could therefore give us more insight on differences in the representation of men and women on Wikipedia. An analysis on all individual attributes included in the 'Family' category was performed.

Finally, following somewhat unexpected results from our primary analysis, we conducted a supplementary analysis of the distribution of birth years by gender. Here, we excluded entries with birth years that are below 10 and above 2014, to prevent faulty entries of people who were too young at that point and people born after the Inferred Gender data set (Bamman & Smith, 2014) was collected. We then normalized the frequency of articles by dividing the number of counts during a certain period of 20 years over the total number of counts over the whole data set for each gender.

All analyses were conducted using the open-source statistical software package, R, and the general-purpose programming language, Python.

Results

In line with previous findings, simple descriptive analysis revealed a strong disparity between the relative presence of male and female entries in our final data set. Out of the 631,258 Wikipedia entries considered in our analysis, a whopping 534,967 (85%) were classified as male. Conversely, only 96,257 (15%) of our entries were female.

Preliminary analysis of the relative prevalence of metadata attributes across all Wikipedia biographies in our data set revealed that less than 5% of entries contained values for the "Known For" and "Net-Worth" attributes (see Figure 1). Furthermore, when grouped by gender, we found less than 0.3% difference between male and female biographies on these attributes. As a result, these categories were excluded from further analyses.

Subsequently, we compared the relative prevalence of our metadata attributes of interest across male and female entries respectively (see Figure 2). Surprisingly, we found that the 'Occupation' attribute was nearly twice as prevalent in female entries than it was in males, with 29% of female pages containing the attributes included in this category, against a mere 19% of male entries. The 'Family' category revealed a

similar discrepancy, with 15% of female entries containing the meta-data attributes within this category, and only 7% of males. For the education category, we observed only a marginal difference in prevalence, with 10% of female versus 9% of male entries containing attributes in this category.

For our unaggregated analysis of the category 'Family', where we compared the relative prevalence of its constituent attributes (i.e., "Children", "Parent", "Relation", "Relative", "Spouse") independently across men and women, we found, most notably, that the "Children" meta-data attribute was over four times as prevalent for female articles than for men. Additionally, whilst relative prevalence of the "Spouse" attribute was low for both men and women, we observed that "Spouse" was reported more than twice as frequently in female entries than in males. Across all remaining family-related sub-attributes, relative prevalence was roughly equal for male and female articles (see Figure 3).

Finally, our supplementary analysis of the distribution of reported birth years by gender revealed a slight, but nonetheless distinct discrepancy between the distributions of birth-years for male and female entries. Whilst the distribution of birth-years was negatively skewed for both male and female entries (with most articles reporting a birth-year within the 20th century), this skew was slightly more evident in females - with proportionally more female entries being born in the last one-hundred years (see Figure 4).

Discussion

In this study, we aimed to quantify the extent to which Wikipedia's biographies represent men and women differently. Overall, we expected that our analysis would yield results consistent with those found in previous studies, hypothesizing that meta-data attributes linked to family relationships would be more prevalent in female entries, whilst categories linked to occupation and education would be more prevalent in males.

However, our analyses revealed a surprisingly mixed set of results. On the one hand, our findings - in line with the literature - indicated that 'Family'-related

meta-data attributes occurred more frequently in female biographies, with children being reported four times more often than they are in male entries. These results offer further support for the notion that Wikipedia's biographies systematically construe women in a way that reinforces traditional gender stereotypes.

On the other hand, however, further results paint a very different picture. Notably, the relative prevalence of both 'Education' and 'Occupation'-related meta-data attributes was found to be higher in women, with the relative prevalence of 'Occupation' being nearly twice as high than it was for male entries. This seems to suggest that, despite an increased likelihood to contain information on children, Wikipedia's female biographies are also more likely to depict their subjects as educated, working professionals - even more so than for males. Taken in isolation, these findings seem to imply the female entries might, conversely, be subject to a "reverse gender bias", where women are more likely to be portrayed in a way that contradicts their traditional gender stereotype.

Yet, how do we reconcile these seemingly contradictory results? At first, we postulated that the increased prevalence of "Occupation" and "Education" meta-data in female entries might be a result of a difference in the distribution of birth years between our male and female sample pools. Here, a relatively "younger" (i.e. relatively more entries from the last century) might, at least in part, account for the fact that education and occupation data occurred more frequently (especially given that women have only begun to gain widespread access to higher education and the professional workforce over the last century (Gaskell & McLaren, 1991)). However, the rather ambiguous results of our supplementary analysis make this a difficult claim to substantiate. Whilst we found birth years of female entries are skewed a *bit* more towards the 20th century than those of the male entries, we do not believe that the difference is strong enough to account completely for the difference in relative prevalence we observe in our main analysis.

Alternatively, these differences could be attributed to efforts by activist groups to improve the quality and completeness of female biographies available on Wikipedia. Women have been an underrepresented group on Wikipedia since the origin of

Wikipedia (Bamman & Smith, 2014; Wagner et al., 2015). Only recently an effort has been made to increase and improve this representation by organizations such as Women in Red (Women-in-Red, n.d.). While the impact of this effort on the total male/female ration on the platform has only been marginal, we may speculate that the effort to improve existing female entries may underlie our observation that female articles are more complete across the board then male ones. Thus, our finding that female biographies more often contain information across all categories than male pages do, could be explained through this recent development.

A final alternative comes from the speculation that women may simply have higher threshold to be included on Wikipedia in general. This would align with the general latency for access of women into educational and professional platforms (Gaskell & McLaren, 1991) and could account for the finding that female biographies seem to be more complete, pointing towards an implicit gender bias in the database itself.

Conclusion

In conclusion, our hypothesis that male biographies will more frequently contain information on education and occupation, whereas female entries will more frequently feature information on family and relationships, could not be confirmed. While our analysis did reveal some discrepancies between the two groups, the limitations of our current data set simply do not allow us to conclusively describe the extent to which these differences are truly underlain by systematic gender bias.

Future research would benefit from using or creating more complete, updated data sets than those currently offered by DBpedia. This may be achieved by conducting a more nuanced linguistic analysis of the attributes utilised in the current study. Such an analysis could include the content of meta-data attributes themselves (as opposed to simply their presence or absence), differences in article length, connectivity between entries, and gender-specific language. Finally, we invite researchers to expand their inclusion of gender to account for those biographies that do not fall into the binary male/female distinction.

References

Bamman, D., & Smith, N. A. (2014). Unsupervised discovery of biographical structure from text. *Transactions of the Association for Computational Linguistics*, *2*, 363–376.

Dutta, A., Roy, R., & Seetharaman, P. (2008). Wikipedia usage patterns: The dynamics of growth. *ICIS 2008 Proceedings*, 172.

Gaskell, J. S., & McLaren, A. T. (1991). *Women and education*. ERIC.

Graells-Garrido, E., Lalmas, M., & Menczer, F. (2015). First women, second sex: Gender bias in wikipedia. *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, 165–174.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al. (2015). Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, *6*(2), 167–195.

Singer, P., Lemmerich, F., West, R., Zia, L., Wulczyn, E., Strohmaier, M., & Leskovec, J. (2017). Why we read wikipedia. *Proceedings of the 26th International Conference on World Wide Web*, 1591–1600.

Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a man's wikipedia? assessing gender inequality in an online encyclopedia. *ICWSM*, 454–463.

Women-in-Red. (n.d.). Wikipedia: Wikiproject women in red. https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red

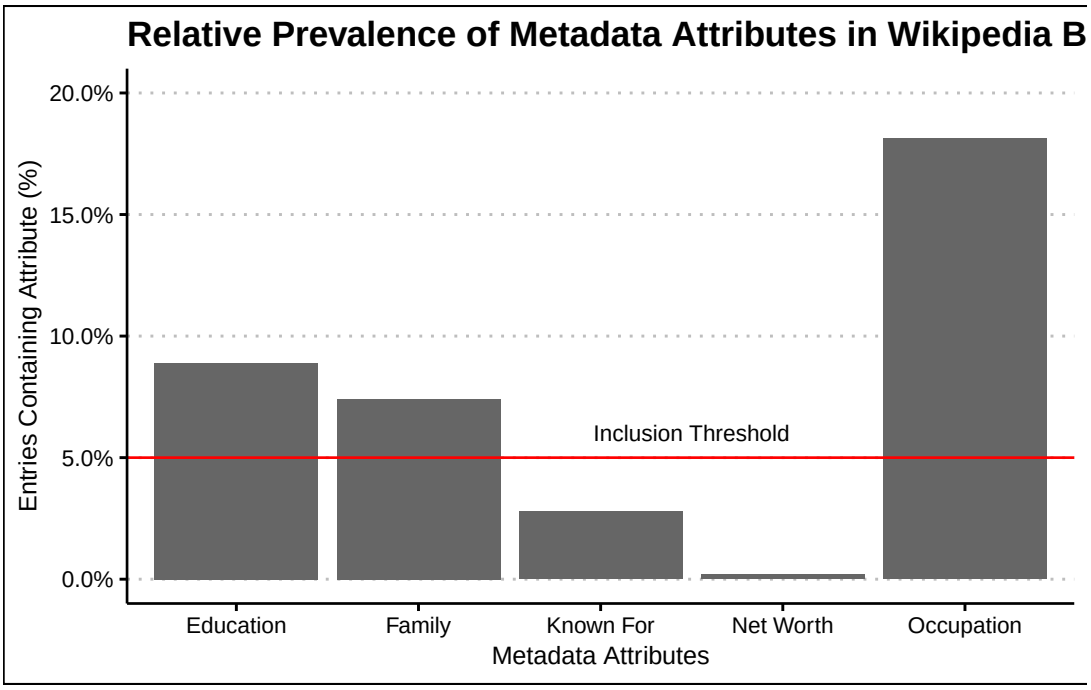

Figure 1. A bar chart illustrating the relative prevalence of information referring to education, family, occupation, "Known For", and net worth across all entries in our data set. An inclusion threshold was arbitrarily set at 5% whereby we include only the attributes that have prevalence higher than that threshold (i.e. Education, Family, and Occupation) in our further analysis

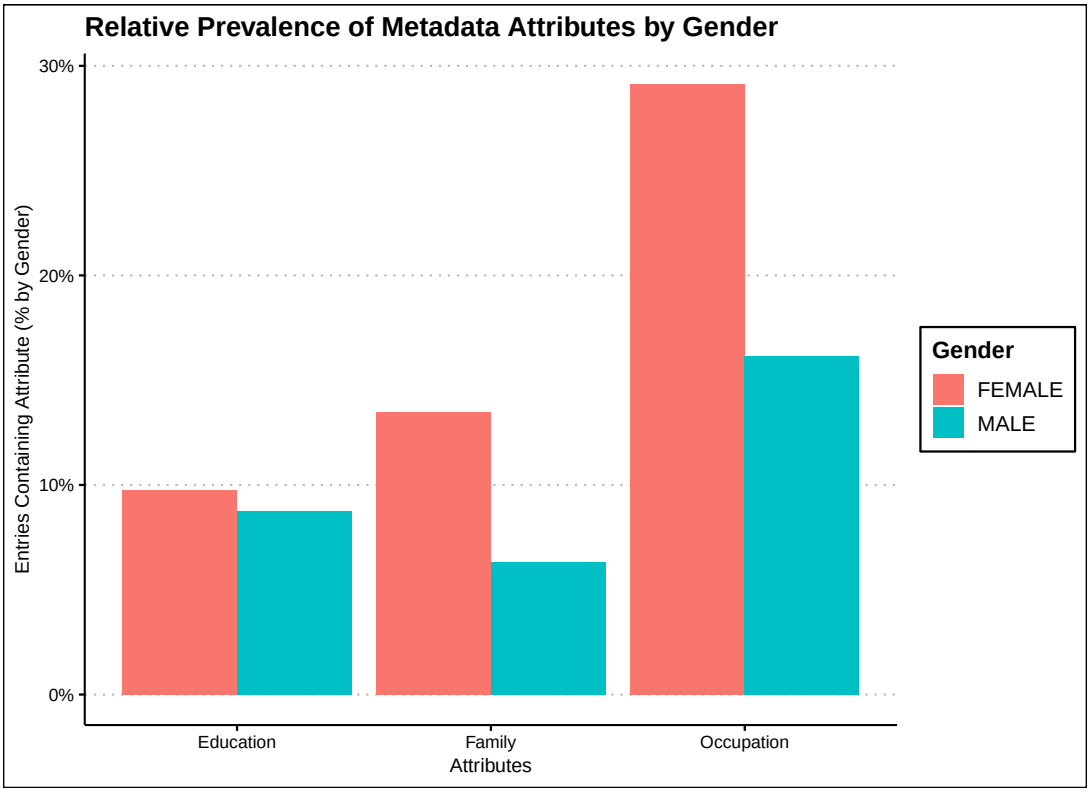

Figure 2. A grouped bar-chart displaying the relative prevalence of three categories containing metadata attributes (i.e., Education, Family, and Occupation) by gender.

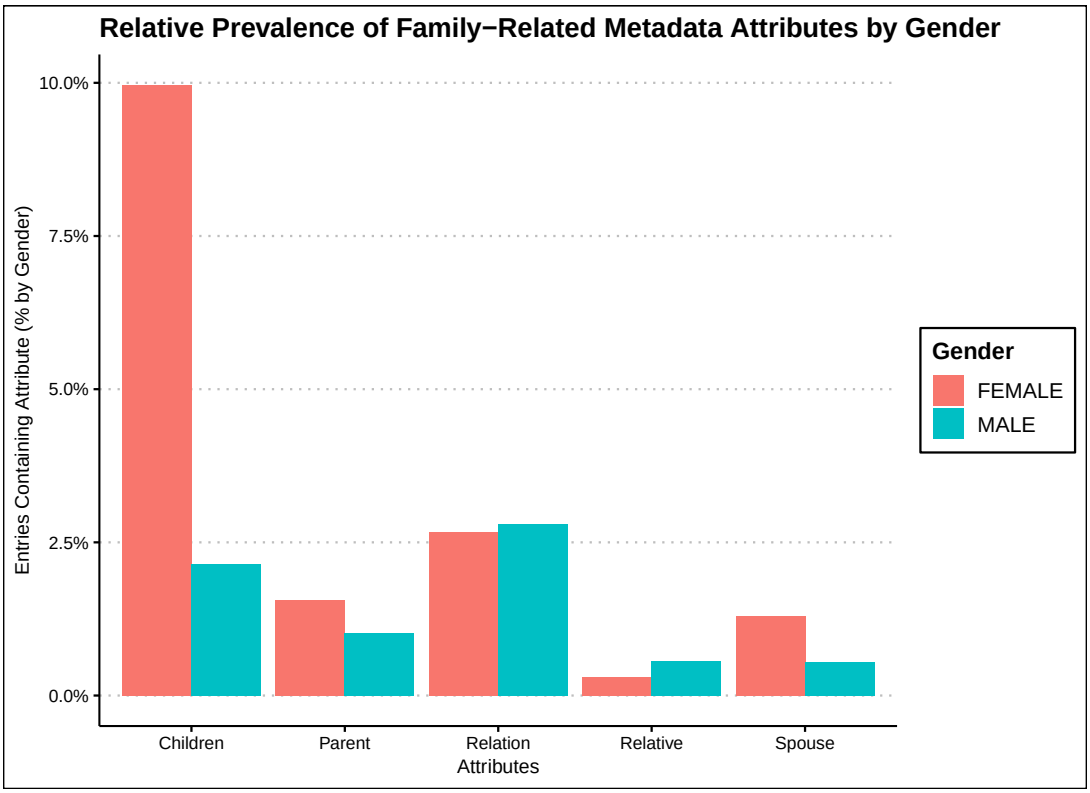

Figure 3. A grouped bar-chart displaying the relative prevalence of unaggregated family-related meta-data attributes (i.e., Children, Parent, Relation, Relative, and Spouse) by gender.

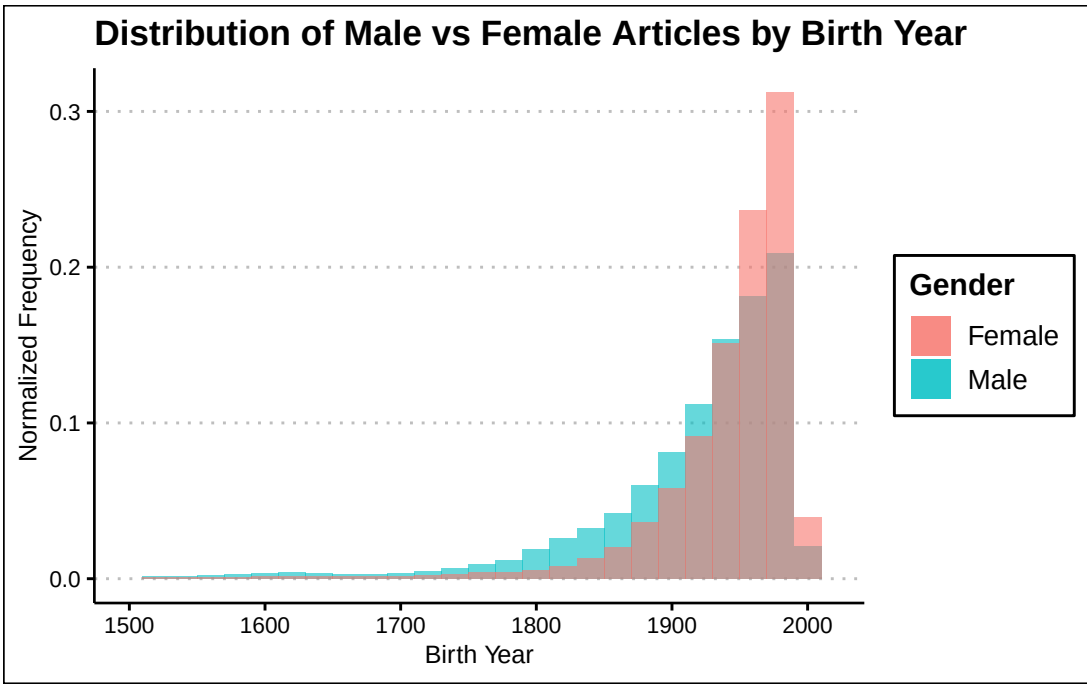

Figure 4. Two histograms representing the normalized distribution of reported birth year for male and female Wikipedia entries.