The Signpost

← BACK TO CONTENTS 18 July 2025 VIEW LATEST ISSUE

Opinion

Women are somewhat under-represented on the English-language Wikipedia, and other observations from analysis

By Yaron Koren Contribute — Share this [show]

The famous gender gap

Roughly 20% of the biographical articles on the English-language Wikipedia are about women - and on seemingly every other Wikipedia, the ratio between male and female article subjects is at least as lopsided. That fact, coupled with the possibly related fact that the percentage of female editors of Wikipedia is roughly 10-15% - has been the subject of numerous news articles, studies, talks, initiatives, and conferences. There are at least 15 different groups that aim to decrease both forms of "gender gap", whether as a primary or secondary goal.

As far as I know, it's rarely stated what the ideal percentage of female subjects should be for Wikipedia articles, though it's at least implied to be 50%. The "Gender gap" hub page on Meta-Wiki essentially states this explicitly, saying that the fact that "more men than women are covered in the mainspace content of our wikis" is a problem that does "harm to the Wikimedia world". But is a 50% ratio actually possible, or even desirable? To believe that may involve believing that, for every man who has achieved notability based on the criteria of the various Wikipedias, there is a woman out there in history with the same notability; so that, presumably, for every Julius Caesar, Mozart or

Thomas Edison, there is a woman of roughly equal historical importance. But human history itself has not been equal, nor has it been fair.

Of course, the average person who is the subject of a Wikipedia article does not have nearly the importance of those three. After all, most professions of the 21st century have a more even gender balance than Roman emperors did. So perhaps the gender gap can be made up for further along the ranks of notability, down with us normal people. This does raise an interesting question: perhaps the ideal gender ratio is not simply a fixed number, but instead a function of the strictness of notability criteria? If anyone were tasked with coming up with a list of, say, the 50 most historically influential people to have ever lived, for example, presumably no one but the most hardline egalitarian would try to include *precisely 25* women. On the other hand, going in the other extreme, an encyclopedia that tried to list every person who has ever lived (currently estimated at 117 billion people), *i.e.* having no notability filter whatsoever, would, if successful, end up with an almost exact gender balance.

Analyzing the notable subject lists

All of this sort of discussion might remain at the level of hand-waving and philosophizing, but there actually are ways to bring some real analysis to this discussion. Extremely helpful here are two different Wikipedia-based initiatives that have attempted to create lists of the most important subjects to cover: "List of articles every Wikipedia should have" and "Vital articles in Wikipedia". Both of these are actually multiple lists: "List of articles every Wikipedia should have" holds 1,000 subjects, while its spinoff listing, "List of articles every Wikipedia should have/Expanded" holds roughly 10,000 subjects. Meanwhile, "Vital articles in Wikipedia" is a set of 5 lists, each one a different "level": the level 1 list holds 10 articles, the level 2 list holds 100 articles, the level 3 list holds 1,000 articles, level 4 holds 10,000 articles, and level 5 holds 50,000 articles.

The level 1 and level 2 listings for "Vital articles" hold no individual people, so that leaves a total of five lists, all carefully curated and maintained, which attempt to contain the most important topics — including the most important people who have ever lived. The careful curation is important, because, looking through the lists, it's hard to dismiss these lists as motivated by any specific political or geographic bias; these lists really do seem to represent an impressive — and dare I say successful — effort to come up with something like a reasonable arbiter of ultimate notability. Even the clichéd white male pop culture enthusiast who prefers to edit the Wikipedia article on, say, Tom Cruise rather than on Juana Inés de la Cruz will presumably have no negative impact on these lists.

In addition to their quality, the other important aspect of these lists is their diversity of size: the fact that they range in length from 1,000 to 50,000 subjects means that we may able to spot how the demographics of the individual humans within the group change as notability standards are relaxed — which may point toward trends that we can extrapolate from.

Methodology

I wrote two PHP scripts that help to analyze all this data. First is a script that scrapes each of these lists, finds the Wikidata entry for the people in that list, and then finds the "sex or gender " value for

that entry - and then generates a CSV file containing all of this data for that list. The second is a script that reads any of these CSV files, and finds the gender breakdown for that list. Both of these scripts (and all of the resulting CSV files) can be found in this GitHub repository , so people can run their own analyses, or find room for improvment in this analysis.

Another note on methodology: there are individuals who are labelled on Wikidata with a gender other than male or female, such as transgender people. The "List of articles every Wikipedia should have/Expanded" list includes one person who does not fall into the two main groups (Judith Butler), while the "Level 5" Vital articles list includes around 20. There is an argument for including all of these in the "female" category, since the gender gap has been described as including them as well; and there is also an argument for having a third category for them. However, ultimately I decided to simply exclude them from the analysis, for the sake of simplicity, and since the relatively small number of such articles (roughly 0.1% of any of the lists) means that their inclusion would not have a significant effect on the numbers in any case.

Results, followed by some extrapolation

Here, then, are the results of this analysis:

List name	Number of articles	Number of people	Number of women	% female
Articles every Wikipedia should have	1,000	203	11	5.4%
Articles every Wikipedia should have, expanded	10,000	1,919	189	9.8%
Vital articles, level 3	1,000	110	9	8.2%
Vital articles, level 4	10,000	1,955	200	10.2%
Vital articles, level 5	50,000	14,645	2,463	16.8%

We can certainly see the trend here: as the notability criteria are broadened, the female percentage rises.

If this basic conclusion is true, then one can imagine putting together a table like this, also taking into account the 117 billion figure for all of humanity:

Number of biographical articles	Fraction of total humanity	Ideal % of female article subjects
110	10^{-9}	8.2%
203	10^{-9}	5.4%
1,919	10^{-8}	9.8%
1,955	10^{-8}	10%

14,645	10^{-7}	16.8%
...		
117 billion	1	50%

What does "Ideal" mean in the table? It means that, if a certain language Wikipedia contains X articles about individual people, and those articles in fact cover the most noteworthy X people of all time, then that is the expected percentage of those articles that will be about women.

Do we dare fill in the rest of the table? It's all rather pseudo-scientific, but the basic premise does seem to make sense. Throwing caution to the wind, perhaps the full table would look something like this:

Number of biographical articles	Fraction of total humanity	Ideal % of female article subjects
110	10^{-9}	8%
203	10^{-9}	5%
1,919	10^{-8}	10%
1,955	10^{-8}	10%
14,645	10^{-7}	17%
117,000	10^{-6}	20%
1.17 million	10^{-5}	25%
11.7 million	0.01%	30%
117 million	0.1%	35%
1.17 billion	1%	40%
11.7 billion	10%	45%
117 billion	1	50%

The known numbers do fit rather nicely into the overall series.

Conclusion

The English-language Wikipedia currently holds roughly 2 million biographical articles. So, according to the aforementioned table, the English-language Wikipedia should have, very roughly, 25% female representation. So there you have it: women indeed are underrepresented on the English Wikipedia — they are 19% of all biographical articles, whereas they should be a little over 25%.

For all other Wikipedias, the ideal fraction will of course be lower. The majority of Wikipedias have fewer than 12,000 articles, which presumably means fewer than 2,000 biographical articles. These Wikipedias, according to the graph, ought to have have 10% of their biographical articles be about

women. (Arguably, we know exactly *which* articles they should have, although that is a more controversial assertion.)

By the way, this kind of analysis could also be done on other demographic traits, like ethnicity, nationality and occupation. By far the easiest trait to do an analysis on, other than gender, though, is year of birth, since data about it is generally comprehensive and uncontroversial. I actually included year of birth in these scripts' output — I did not mention it so far in this essay because the subject gets a lot less discussion than the gender ratio, though it does show up in discussions of "recentism". But the results for birth year are, interestingly, even more dramatic than for gender. One startling finding is that, in the "Level 5 Vital Articles" list, people born in 1922 or later make up a full 39% of the list; while the much smaller "Level 3" list holds only one person born after 1922 (Michael Jackson), and thus less than 1% of the overall list. In the intermediate "Level 4" list, the number is in the middle, at 17%; so again we can see this sort of logarithmic progression.

This type of analysis could lend itself to all sorts of observations about Wikipedia's deficiencies relating to different demographic groups; more broadly, it could be used to study the historical importance of different areas and groups over time (e.g., how important was 15th century Italy?).

As they say, further research is warranted.

← *PREVIOUS "Opinion"*

DISCUSS THIS STORY

+ Add a comment

THESE COMMENTS ARE AUTOMATICALLY TRANSCLUDED FROM THIS ARTICLE'S TALK PAGE. TO FOLLOW COMMENTS, ADD THE PAGE TO YOUR WATCHLIST. IF YOUR COMMENT HAS NOT APPEARED HERE, YOU CAN TRY PURGING THE CACHE.

- "Even the clichéd white male pop culture enthusiast who prefers to edit the Wikipedia article on, say, Tom Cruise rather than on Juana Inés de la Cruz will presumably have no negative impact on these lists." Quite a presumption. Innisfree987 (talk) 08:39, 18 July 2025 (UTC) [reply]

IN THIS ISSUE

- Comix
- **Opinion**
- Community view
- Obituary
- Traffic report
- Humour

IT'S YOUR *SIGNPOST*. YOU CAN HELP US.

Categories: Wikipedia Signpost archives 2025-07 | Wikipedia Signpost RSS feed

---