Women's presence on Wikipedia: A comprehensive analysis on the evolution of women's recognition over time, country and fields of work.

Emile Bourban
emile.bourban@epfl.ch

Florian Delberghe
florian.delberghe@epfl.ch

Abstract

In last century, in most countries, women have finally gained more rights and we have been continuously progressing towards a more equal society. However, their representation and recognition in media hasn't always been fair. In this study, we wanted to investigate how Wikipedia biographies can help shed light on women's achievements across time, culture and different fields of work.

1 Introduction

It is a well documented fact that encyclopedias in general have always had some gender bias; the same can be said for Wikipedia. Even though there are no direct restrictions on who can create or edit, the articles they are mostly written by men (Zurn, 2011). The aim of this study is solely focused on the articles and not the editors. We have previously hypothesized that due to the closing gender gap nowadays, we should start to see a form of gender equilibrium being reached. For this analysis, we will focus on two major topics, the differences in women representation between different cultures and different fields of work. We started the project using pre-processed information about Wikidata from WHGI, then, seeing that it lacked some of the required details, we chose to extend our dataset by parsing the entire Wikidata JSON file.

2 Related Work

Some work was already done on the subject that studies the inequalities between genders through time and geography (Konieczky and Klein, 2018). Their conclusion was that gender disparities is a phenomenon with a long history but that also presents some longitudinal component related to culture differences. The analysis using the Inglehart–Welzel cultural clusters has inspired us to do our own clustering among cultures to see different trends.

3 First Look at the Data Using WHGI

3.1 The data

The data used for the first part of this project was gathered from (WHGI, 2018). WHGI is an open data project producing *.csv* files regrouped by gender, occupation, dates and places of birth, etc, from as many Wikipedia pages as possible. This data is quite user-friendly and was thus our first point of approach in this study.

3.2 Changes Through Time

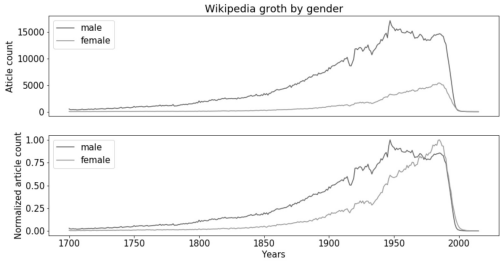

Figure 1: Evolution of the number of Wikipedia articles by gender.
Full count (top) and normalized (bottom)

The first thing that we wanted to show was the evolution of the number of articles per gender through time, using the birth dates to estimate the period the person belonged to. In Fig 1 is shown the evolution of the article count for both genders. We can observe that throughout time the article count for men has always been higher than for women. The later part of the graph shows a slowing of the growth of articles about men but women still have to this day an increasing presence on Wikipedia – especially visible on the normalized graph –. The combined decrease for men and in-

crease for women could mean a possible equilibrium in a few decades.

3.3 Differences Across Cultures

We have found great homogeneity when comparing ethnic or geographic group. It seems that across cultures (taking only the largest ones to have significant data) the ratio of women to men is the same. There are however some remaining outliers; african-americans have over represented women and a large proportion of the people on wikipedia that are born in Tokyo are women. This peculiarity of Japan will be further explored later using more data.

4 Data Extraction

After of bit of analysis, we came to the conclusion that the size and complexity of the data provided by the WHGI was too small. We then decided to introduce a way to parse and utilize the whole Wikidata JSON dump to scrap more information.

4.1 The Wikidata Database

Wikidata is a public and collectively edited dataset hosted by the Wikipedia fundation. They aim at providing a free common source of data (Wikidata, 2018)

4.1.1 Description of the Dataset

The data was downloaded in the form of a compressed JSON file. It contains information about all the Wikipedia articles regrouped in the same format. Each article is assigned a label (title of the article) and a unique item identifier. Inside of every instance, we can find many properties (in the PXXXX format) that references a unique characteristic of the person; it can be the date of birth, citizenship or any other values that can be assigned to a specific article. Each of those property can take different values collected in the same instance and stored as Wikidata item identifiers which looks like: QXXXXX.

4.2 Fields of Interest

For this analysis, we have decided to focus on those 8 fields: - Birth date - Place of birth - Citizenship - Date of death - Gender - Id - Name - Occupation - Article language - As they provide the most important information for our later analysis.

4.3 Data Cleaning and Treatment

The previously acquired data is raw and thus contains a lot of erroneous values and unusable fields. The first step consists in extracting the date from the datetime string and storing only the years of death and birth (this gives us enough information already); since the datetime format can be poorly handled by spark, this value was stored as an integer. The genders were also extracted form the Wikidata Id codes and set to 'Male', 'Female' or None. Other genders were not taken into account since their number is insignificant ant it was not the aim of our inquiries.

Without date and gender information the row was considered meaningless and dropped. We were only interested in recent data in order to have enough of it so we chose to only select dates after 1700. We also noticed some wrong values from the thai Wikipedia that are set in the future. After inspection they do not have any usable data attached to them and were discarded as well.

Another quirk of the data is the fact that when the precise date of birth or death is missing the value is set to the 1st of January of the same year. Some pages that were missing dates altogether and were assigned to a random year causing some large peaks that were ignored as well.

4.4 Grouping and Choices for Cultural Clusters

The raw data is composed of many citizenships and languages some of which can be very sparse. For this caveat we chose to inspire ourselves with the cultural clustering done in (Konieczky and Klein, 2018) and regroup the nationalities by ethnic or geographic groups. We have chosen several groups that can be seen in our *regions dictionary* in the notebook, those were made from apparent similarities between cultures.

5 Effects of Different Cultures on the Recognition of Women

5.1 Similar Cultures Show Similar Evolution

The first thing we wanted to do was the comparative analysis of similar cultures; in our case, this meant using the data for western europe and north america. When displaying the article count, we can see that even though the counts are different the shapes of the normalized curves are very similar. This observation prouves our previous hypothesis that similar cultures - even if geographicaly

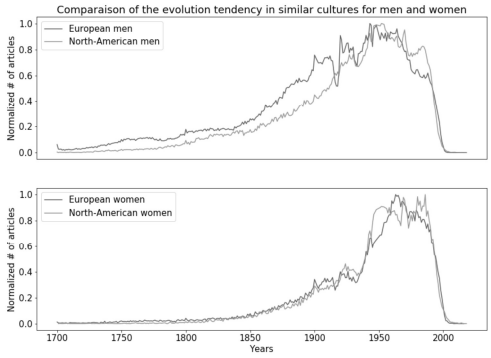

Figure 2: Comparative Evolution in North-America and Europe

distant - show similar trends. Another interesting point is to look at the dips on the curves during the periods of World War I and II. For the first one, there is a decrease in the values in the european curve but no consequences on the american trend. Whereas for the second World War, similar decreases show on both curves. This interesting observation also validates the hypothesis that the Wikidata follows the actual population trends for each country.

5.2 The North East Asian Example

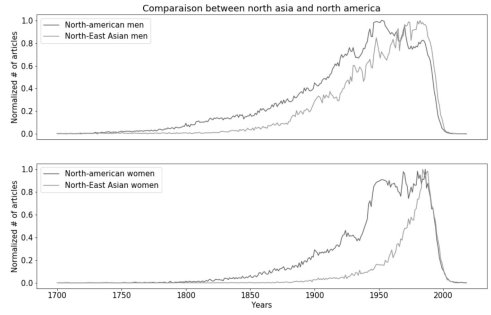

Figure 3: Evolution in Asia Compared to North-America
Shows normalized evolution to compare the different evolution of women's recognition in asia

For this part of the analysis, we were inspired by the fact that there are a lot of women on the Japanese Wikipedia. Comparing the article count, we can see that today, for this geographical group the equality between genders has been reached. This has not always been the case however. By comparing it to the evolution in north-america we can see a surprising disparity. The shape for the men is quite similar but for the women, the curve starts to go up later in time and its increase is exponential up until it reaches about the same amount of articles as for the men. This is the only location that we studied that has that high a level of equality but the way it gets to this point is very singular.

5.3 Arabic Coutries and Women

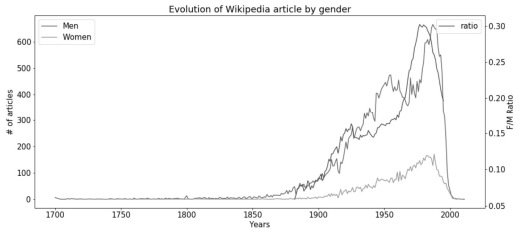

Figure 4: Progression of Gender in Muslim Countries
Article count and ratio for Middle-Eastern and Mahgreb coutries

In this section, the idea was to show whether arabic majority countries had a bias toward women. We hypothesized that in those parts of the world where women have less rights (Women's suffrage, 2018) we could see a lower proportion of women on Wikipedia. We can observe that the start of the shape is similar to the ones seen previously. Then when comparing the M/F ratio the maximum value reached is 0.30 whereas for the western Wikipedia, the max ratio is around 0.40. One other observation is that the max ratio is reached in the 70s as compared to the end of the 90s for other regions. This may be due to the fact that during this period some of the middle eastern countries were a lot more secular than they are today. Then, a more recent rise of religiosity has caused increased disparities that may explain this behavior.

6 Women's Presence Across Domains

After regrouping the people on Wikidata by domain of occupation, we were able to get a picture of the evolution of gender equality in different domains. In Fig. 5 we observe that for most domains, the ratio of women increases over the years which seems to reflect well the situation worldwide were most occupations tend towards more and more equality. We can see however that most occupations still do not have equal ratio of men and women: only literature and journalism and social science reach a ratio that is close to 0.5 and

entertainment is the only occupation where the ratio is noticeably higher than 0.5. Most categories are still below a 0.5 ratio but the trends looks to be going the right way for all of them. Even in categories that were exclusively masculine until recently like religion or military and law enforcement, the ratio is going up steadily.

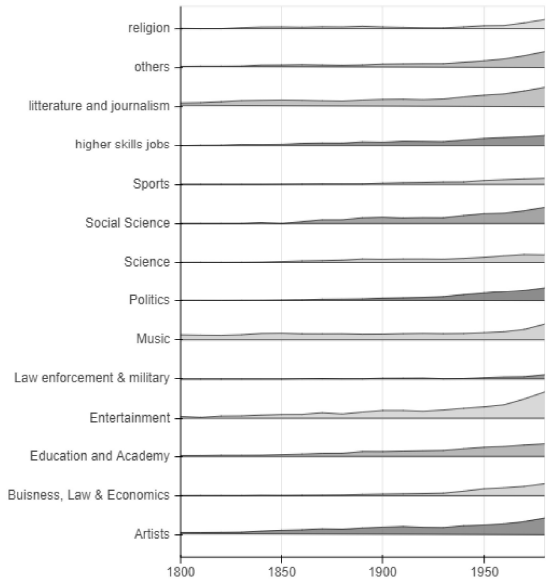

Figure 5: Evolution of ration of women through time across different domains of occupations.
The distance between each line corresponds to a ratio of 1

7 Women's Presence Through Widipedia Article Language

On Fig. 6 we can see that the ratio of women seems to be fairly homogeneous across Wikipedia languages. The outliers that we can observe here are only for rare languages and dialects that have fewer articles, and thus they might be the result of random variability. Furthermore, they are very heterogenous geographically and culturally which leads us to conclude that the languages of the articles in Wikipedia can not be used to represent the gender disparities.

8 Conclusion

Through these analysis, we can conclude that Wikidata can indeed be used to highlight gender inequalities across time, culture or space. We have shown that those inequalities are reflected in the representation of women on Wikipedia and that the difference that we expected between different culture does indeed appear. We also showed that though equality has different levels in different

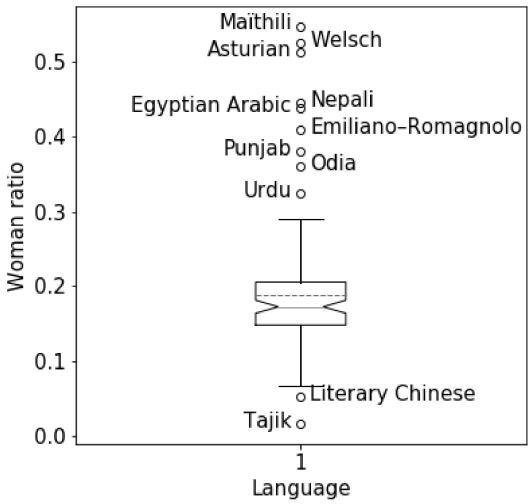

Figure 6: Boxplot of distribution of ratio of women for different Wikipedia article languages.
The median is shown in red with a notched 95% confidence interval and the mean is represented by the green dotted line

places and does not seem to have been reached yet, the trends point toward an ever increasing equality, in all domains of occupations or countries.

There are however some limitations to our approach. Firstly, we know that Wikipedia only features people that are famous and or have accomplished something significant which is not necessarily representative of every aspect of society. For instance, wage-gaps could still be present even with an equal distribution of gender for famous people.

Secondly, the approach we used here is limited in its temporal resolution. Indeed, since the vast majority of people on Wikipedia obtain their fame in their adult life, it is impossible to have an accurate representation of the current time and we can only make some analysis up to about 20 - 30 years in the past.

Thirdly, though we have enough data to make some global representation, the data often becomes sparse when we restrict ourselves to a specific subset in order to analyze precise places or cultures, and the random variability makes it harder to extract significant results in these cases.

To conclude, we can say that even with these limitations, the equality between men and women is shown to still be an issue in most of the world, but the trends that we observe leave us hopeful that we will reach it in a not so distant future.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Rhonda Zurn. 2011. *Gender gap shows no sign of closing over the past five years*

Piotr Konieczny. Maximilian Klein. 2018. *Gender gap through time and space: A journey through Wikipedia biographies via the Wikidata Human Gender Indicator*

Wikidata Home Page 2018. *url: https://www.wikidata.org/wiki/Wikidata:Main_Page.*

Wikidata Human Gender Indicators (WHGI) 2018. *url: http://whgi.wmflabs.org/*

Timeline of women's suffrage 2018. *url: https://en.wikipedia.org/wiki/Timeline*$_o f_w omen\%27s\_suffrage$