

# Visual Gender Biases in Wikipedia: A Systematic Evaluation across the Ten Most Spoken Languages

Pablo Beytía,<sup>1</sup> Pushkal Agarwal,<sup>2</sup> Miriam Redi,<sup>2,3</sup> Vivek K. Singh<sup>4,5</sup>

<sup>1</sup>Humboldt University of Berlin, <sup>2</sup>King’s College London, <sup>3</sup>Wikimedia Foundation,

<sup>4</sup>Rutgers University, <sup>5</sup>Massachusetts Institute of Technology

beytiapa@hu-berlin.de, pushkal.agarwal@kcl.ac.uk, miriam@wikimedia.org, v.singh@rutgers.edu

## Abstract

Wikipedia, a collaborative crowd-sourced platform, is the largest, most visited online encyclopedia. Since it spreads information freely in more than 300 languages, many users, tools, and dashboards rely on its content. Hence, there is a need to maintain its fairness and completeness. However, previous research has indicated the existence of a significant gender gap in Wikipedia biographical articles. We already know that a minimal proportion of those articles portray women and there are gender asymmetries in the textual content of these articles, but little has been reported about the *visual* aspects (e.g., image volume or image quality) of the gender gap. Here, we analyze all biographies available on Wikipedia across 300 occupations in the ten most widely spoken languages, and undertake quantitative and qualitative analysis of gender differences in the written and visual content. The cross-lingual results indicate that (1) much of the male bias in content arises when editors select which personalities should have a Wikipedia page, (2) the trends in written and visual content are quite dissimilar, (3) men biographies tend to have more images across languages, and (4) female biographies average better visual quality. A more granular analysis is performed on English Wikipedia, distinguishing trends in occupations and qualitatively analyzing science and technology biographies. The overall results shed light on the kinds of visual biases that emerge in the collaborative creation of Wikipedia and yield guidelines for future management of contributions on the platform.

## 1 Introduction

In recent years, a broad scientific interest in Wikipedia’s knowledge gaps has emerged. Several studies have pointed out that “the encyclopedia that everyone can edit” tends to register systematic disparities in various content dimensions (Redi et al. 2021). Researchers have identified significant inequities concerning coverage of times, geographic locations, cultures, genders, languages, and socio-economic circumstances (Graham et al. 2014; Gruwell 2015; Samoilenko et al. 2017; Beytía 2020; Agarwal et al. 2020).

In this context, the topic of *gender asymmetries in content* has received quite a lot of attention. Across all its language



हिन्द की बेटियाँ किसी से कम नहीं हैं।

“हिन्द की बेटियाँ” परियोजना में योगदान के लिये आपकी जितनी प्रशंसा की जाय कम है।  
-- मुजम्मिल (वार्ता) 13:39, 26 नवम्बर 2013 (UTC).

Figure 1: Wikiproject “Daughter’s of Hind” portraying Queen of Jhansi–Laxmi Bai, an initiative for “Empowering Indian women” on Wikipedia<sup>1</sup>. The text in Hindi translates as – Daughters of Hind are not less than anyone. We are really thankful for your help on this project.

editions, Wikipedia currently stores more than 6M biographies about “notable” people. Of these, less than 20% portray women (Wikimedia 2020)<sup>2</sup>. Continued exposure to such biased content could foster and fortify gender prejudices and stereotypes globally (Wajcman 2008) and widen the gender gap in Wikipedia audiences (Johnson and Williams 2020).

Numerous initiatives have emerged over the years to close this gender gap. For example several projects were created to expand women’s coverage on Wikipedia in English and other languages: Women in Red, Editatona, Wiki Loves Women, WikiGap, Wikiproject “Daughter’s of Hind” (see Figure 1), among others<sup>3</sup>. These projects allow for addition of images for women, from different parts of the world, however vital gaps still remain.

In recent years, researchers have also contributed to this cause by studying and measuring content gender gaps. They have reported that the women portrayed in Wikipedia are on average slightly more notable than men and due to a ‘*glass ceiling effect*’: only very notable women get to be included in this registry, while men would have lower barriers to entry (Wagner et al. 2016). Moreover, studies have shown that the content gender gap is a complex phenomenon composed of several overlapping information asymmetries. Empirical evidence reveals that inequity in *coverage* is complemented by imbalances in articles’ *topics*, *lexicon*, *classification*, and *structural position* (Graells-Garrido, Lalmas, and Menczer

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org).

<sup>1</sup>bit.ly/Lakshmibai

<sup>2</sup>Data extracted in February 2020.

<sup>3</sup>Some of the groups and initiatives are listed here: [https://meta.wikimedia.org/wiki/Gender\\_gap](https://meta.wikimedia.org/wiki/Gender_gap)

2015; Wagner et al. 2015; 2016; Beytía and Wagner 2022). Lastly, ‘Wikidata Human Gender Indicators’ (Klein et al. 2016) delivers weekly updated statistics on (some facets of) gender disparities on Wikipedia.

However, there exists much less work regarding *visual* content biases on Wikipedia. Despite the crucial role of images for gender roles and stereotype diffusion (Goffman 1979), they have been mostly overlooked in research on gender asymmetries on Wikipedia. The few existing studies analyzing visual disparities across genders (Young, Wigdor, and Kane 2016; Singh et al. 2020; Zagovora, Flöck, and Wagner 2017) are relatively small scale in terms of quantity of biographies analyzed (1,000 articles at most), languages considered (English or German only), and range of metrics.

This study aims to overcome these limitations to advance the development of an overall, systematic perspective on visual gender biases in Wikipedia. For the first time, we conduct a macro cross-lingual exploration of gender gaps in Wikipedia biographies from a multimodal perspective. In doing so, we make multiple contributions:

- We collect a dataset of *all* 6.2M biographical articles available on Wikipedia and the images included in biographies for *the ten most widely spoken languages* in the world<sup>4</sup> (English, Mandarin Chinese, Hindi, Spanish, French, Standard Arabic, Bengali, Russian, Portuguese, and Indonesian).
- We define and compute 4 multimodal metrics to measure and compare the gender gap across the biographies’ *visual* and *textual* components. These metrics evaluate content *quantity and quality*, and cover the three ‘stages of visibility production’ in Wikipedia – selection, building, and positioning – as proposed by Beytía and Wagner (2022).
- We explore *monolingual and multilingual* gender asymmetries across ten occupational categories that cover the 300 most typical occupations on Wikipedia.

The more comprehensive results from the analysis suggest that: (1) much of the male bias emerges at the article *selection* stage (i.e., deciding which personalities should have a Wikipedia page), (2) the trends in visual and textual biases are dissimilar, (3) men biographies tend to have more images across languages, and (4) female biographies average better visual quality.

The next Section reviews the scientific literature on the Wikipedia content gender gap. Subsequently, we explain our data collection methodology and describe the metrics we use to measure content gender biases (Section 3). Section 4 visualizes the main quantitative results and qualitatively analyzes the gender disparity in biographies of science and technology. Finally, we discuss our overall findings (Section 5) and present conclusions (Section 6).

## 2 Background and Related Work

Collaborative editorial work on Wikipedia inevitably produces patterns of visibility for genders (Beytía and Wag-

ner 2022). It illuminates each one in a particular way, giving it a degree of *representation* (e.g., number of articles), a form of *characterization* (e.g., associating its articles with certain topics), and a *structural placement* in the information system (e.g., disseminating its content in a number of languages). Following this idea, we can classify the literature on gender asymmetries in Wikipedia content into three ‘stages of visibility production’: selection, building, and positioning (Beytía and Wagner 2022).

The first stage involves *content selection* processes, including the suggestion of biographies to be developed in articles and the dynamics of acceptance or deletion of previously elaborated entries. At this phase, what is at stake is the degree of *representation* of each gender in the biographies repository. Several investigations have evaluated genders’ *coverage* in Wikipedia, indicating that women have fewer biographies than men, representing between 13.2% and 19.9% of the total record (Graells-Garrido, Lalmas, and Menczer 2015; Yu et al. 2016; Wikimedia 2020). By contrast, no systematic gender differences have been found in the content *removal* processes. Studies suggest that there are no more deletions of articles about women than men and no more requests for deleting content about women (Adams, Brückner, and Naslund 2019; Worku et al. 2020).

The second stage comprises the dynamics of *content building*. It involves the insertion of written and audio-visual information in articles, considering the processes of choosing, structuring, and presenting that content. At stake in this phase is the *characterization* of each gender in Wikipedia’s discourse. Investigations found no discrimination against women in terms of the *length* of biographies. In fact, female biographies tend to be longer than male ones (Graells-Garrido, Lalmas, and Menczer 2015; Wagner et al. 2015). Something similar happens with encyclopedic *references* and *sources*: research on profession-specific articles shows that biographies about women have more references than those about men, and those references come from more diverse sources (Young, Wigdor, and Kane 2016). However, there are systematic asymmetries in the *topics* discussed in the biographies and in the *lexicon* used to write them. Women’s biographies tend to focus more on gender, social relations, and family characteristics (Wagner et al. 2016) and more frequently use words associated with gender, achievement, and family (Graells-Garrido, Lalmas, and Menczer 2015).

The third stage includes *content positioning* processes, considering the association (classification and linking) of articles and the dissemination of them in multiple languages. At this phase, what is at stake is the *structural placement* of the articles of each gender in Wikipedia, i.e., their location in encyclopedic categories and their centrality within the information system. Studies on *classification* show that biographies of men are more frequent in sports categories and those of women in arts categories (Graells-Garrido, Lalmas, and Menczer 2015). In the *hyperlink network* between articles, biographies with higher centrality are predominantly male, and connectivity between articles about men and women is often asymmetric and favors men (Graells-Garrido, Lalmas, and Menczer 2015; Wagner et al. 2015). However, there

<sup>4</sup><https://www.visualcapitalist.com/100-most-spoken-languages/>

is no male dominance concerning *multilingual notability*: women, on average, tend to have their biographies in more languages than men, even controlling for occupation type and year of birth (Wagner et al. 2016).

Despite the cross-cutting nature of the visual field, few studies have analyzed how images on this platform establish gender asymmetries, and the exceptions have focused on specific occupational groups, on particular languages, or have not explored biographical articles.

Young, Wigdor, and Kane (2016) analyzed the biographies of the Fortune 1000 CEOs in English Wikipedia. They concluded that there is no gender bias in this group of articles' visual content, at least when looking at the total number of images they include and the number of images in which each CEO appears. Zagovora, Flöck, and Wagner (2017) examined articles describing occupations (not biographies) in German Wikipedia and found a gender disproportion: almost half of the images (44.8%) show men practicing the professions and only 12.4% depict women. Singh et al. (2020) analyzed in English Wikipedia images of traditionally female professions (Librarian, Nurse) and male professions (Computer Programmer, Civil Engineer), and then compared the gender frequencies in those images with U.S. labor statistics. The results suggest that the images in Wikipedia have a male over-representation: they exacerbate the presence of men in typically male professions and also underrepresent women in typically female professions.

There remain multiple gaps in the current analysis of visual gender asymmetries on Wikipedia. This includes covering all biographical articles, considering all professions, and working across multiple languages. Further a comparison between visual and textual gaps is rarely studied, and lastly, images have only been analyzed quantitatively (and not qualitatively). Hence, this work marks a significant step toward comprehensive understanding of gender bias in visual content on Wikipedia.

### 3 Methods

This Section presents our methodology and dataset details. We collect a large dataset of Wikipedia biographies and their images, and make it publicly available<sup>5</sup>. We develop metrics using biographies' text, images, and additional demographic data related to the articles.

#### 3.1 Methodology to Gather the Dataset

Since the focus of our study is to understand articles on people, we start by collecting a list of 6.22 Million people available from Wikidata as of February 2020. Wikidata<sup>6</sup> is a knowledge graph storing most of the knowledge on Wikipedia and its sister projects in a structured form. For example, the Wikidata ID for Serena Williams is *Q11459*. Using this ID, one can view or fetch machine-readable details from Wikidata in a `property:value` format<sup>7</sup>. We

<sup>5</sup>Dataset: <https://doi.org/10.5281/zenodo.5899657>

<sup>6</sup><https://www.wikidata.org/>

<sup>7</sup>View available at <http://wikidata.org/wiki/Q11459>. For each person, we extract Gender (P21), Occupation type (P106), Birthplace (P19) and the list of Wikipedia articles in different languages

describe the above extracted features as follows:

- *Gender*: For each Wikidata item in our list, we take the gender values from and recode them into three categories: male, female and other, which includes a small list of non-binary gender identities. The resulting gender distribution is as follows: 70% Males, 20% Females, 0.02% Other, and nearly 10% are Unknown (i.e., information is missing).
- *Occupation*: We take the occupation value from the Wikidata items. When the item has multiple occupation values, we take at the top of the list, namely the most relevant and referenced occupation of a person. To standardise the occupations across all biographies we hand-coded the most frequent 300 occupations, which covers 61% data, into 10 broader occupation categories. Note that, of the remaining 39%, around 36% are unknown or null. This coding uses labels from wide spectrum of taxonomy defined in Pantheon project (Yu et al. 2016)<sup>8</sup> and reviewed in the Networked Pantheon database (Beytfa and Schobin 2020). The taxonomy groups various professions to a topical category. For example, all professions like actor, writer, painter, singer, poet and so forth will come under 'Arts' category. The top three occupation in our dataset are Arts, Sports, and Science and Technology, with 16%, 13% and 12% biographies respectively. We show the distribution of items by birthplace and gender in Figure 2.
- *Birthplace Continent*: We take this as an indicator of geographic location associated with person. From Wikidata, we resolve the birthplace location into longitude-latitude coordinates, based on which we obtain country and continent labels. A large proportion of biographies (65%) are missing the locations field. We performed a manual check on a random subset of Geo-tagged data and it showed that nearly all times the Geo-tagging is accurate with our pipeline. Europe is the top continent in our data accounting for nearly 20% of the biographies. We show the distribution of items by occupation and gender in Figure 2.

We additionally check if the missing data in occupation or birthplace continent adds any significant bias in data distribution. We test this by recreating the distribution plots (Figure 2) using the data from the missing biographies and then comparing those distributions with our data set. First, we check gender with continent plot for people not included in our analysis because they do not have one of the top occupations. Then, we replicate the gender with occupation plot for people who have no specified birthplace continent.

We do not find any relevant change in the gender with continents plot rendered with missing biographies (without a top occupation). In the gender with occupation plot (biographies with the missing continent), we find minor changes in the order of the occupations (i.e. the ranking of frequencies). However, the Spearman rank correlation test confirms that the ordered lists of occupations in our sample is highly related with the same list with the missing cases ( $value = 0.93, p < 0.0001$ ).

about that person.

<sup>8</sup><https://www.kaggle.com/mit/pantheon-project>

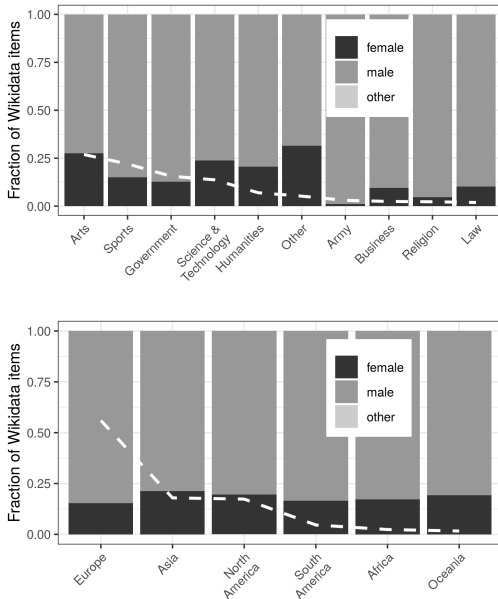


Figure 2: Breakdown of people items on Wikidata by gender plus occupation type (Top) and birth-place continent (Bottom). The white dotted line shows the fraction of items for the category on the x-axis, after removing null values. The x-axis is sorted based on decreasing order of these fractions.

### 3.2 Collecting Article Data

For each item on Wikidata, we collect the list of languages for which there exists a corresponding Wikipedia article. We see that pages have been created on an average in 1.6 different languages for the items in our dataset. The page present in maximum languages is that on Barack Obama (235 languages). Given the wide spectrum of languages available on Wikipedia, and the variability in terms of focus and volume of data, for the purpose of this study, we focus on 10 main languages only. Specifically, we focus on the languages with the most speakers: English, Spanish, French, Portuguese, Mandarin Chinese, Indonesian, Hindi, Bengali, Russian and Arabic. This is because the structuring of knowledge in these languages probably has the greatest consequences in contemporary global culture. We show number of biographies with/without images and images in each language in Table 1.

We also calculate the *article quality* for each biography in the 10 language considered. To do so, we use Wikimedia’s Language Agnostic Quality Classifier<sup>9</sup> that automatically scores Wikipedia article revisions based on their overall quality, using structural information of the article such as the number of references, the length, etc. The output of the classifier is a score in the range  $[0, 1]$ , where 1 corresponds to high quality, and 0 to stub-level article quality<sup>10</sup>.

<sup>9</sup>[meta.wikimedia.org/wiki/Research:Prioritization\\_of\\_Wikipedia\\_Articles/Language-Agnostic\\_Quality](https://meta.wikimedia.org/wiki/Research:Prioritization_of_Wikipedia_Articles/Language-Agnostic_Quality)

<sup>10</sup>[en.wikipedia.org/wiki/Wikipedia:Content\\_assessment](https://en.wikipedia.org/wiki/Wikipedia:Content_assessment)

Language	#Biographies with(out) images	#Images
Arabic	330k (119k)	452k
Bengali	12k (11k)	18k
Chinese	74k (143k)	109k
English	658k (1.02M)	1.04M
French	420k (168k)	743k
Hindi	11k (7k)	16k
Indonesian	47k (29k)	64k
Portuguese	182k (46k)	276k
Russian	266k (166k)	437k
Spanish	197k (226k)	316k

Table 1: Dataset details per language.

### 3.3 Collecting Image Information

After collecting people biographies, we collect the links of all images on these Wikipedia pages using the *image* table from the Mediawiki database<sup>11</sup>. For the 10 selected languages we retrieve a total number of 4M image urls (1.4M unique), out of which we retain 2.86 M (1.37M unique) images and remove image files having .svg, .png, or .gif extension which mostly represent icons and appear in nearly all pages (e.g., Wikipedia logo).

We show the average number of images per page (if they exist) in each language in Figure 3 (Left). Figure 3 (Middle and Right) shows the fraction of biographies covered per language (with and without image) and the relative ratio of biographies with and without images. Note that this is based on selected 10 languages. For each image in our dataset, we also calculate its intrinsic *visual quality* by training a Wikipedia Image Quality classifier. To do so, we collect a training set of 300k images. The set is composed of:

- The positive samples, namely 150k *High quality images*. These are images tagged with the *Quality images* category from Wikimedia Commons<sup>12</sup>. These images are likely to be of high quality, as editors add images to the *Quality images* images category only if they meet specific image quality requirements decided by the Commons community<sup>13</sup>.
- The negative samples, namely 150k *low quality images*. These are images randomly drawn from the repository of Commons images. Random Commons images are likely to be of low quality, small resolution, and they are rarely used to illustrate Wikipedia articles (Erickson, Perez, and Perez 2018). The likelihood that these images are of high quality is very low. At the time of writing, there are around 275K images marked as Quality images on Commons, which is around 0.4% of the images available in the repository (circa 76M).

We next train a classifier able to distinguish low (negative samples) VS. high (positive samples) quality images. To do so, we train an Xception (Chollet 2017) neural network. We initialize the network with random weights, and train for 50 iterations, with a batch size of 32. We use 80% of the

<sup>11</sup>[mediawiki.org/wiki/Category:MediaWiki\\_database\\_tables](https://mediawiki.org/wiki/Category:MediaWiki_database_tables)

<sup>12</sup>[commons.wikimedia.org/wiki/Commons:Quality\\_images](https://commons.wikimedia.org/wiki/Commons:Quality_images)

<sup>13</sup>[commons.wikimedia.org/wiki/Commons:Quality\\_images\\_candidates/#Guidelines](https://commons.wikimedia.org/wiki/Commons:Quality_images_candidates/#Guidelines)

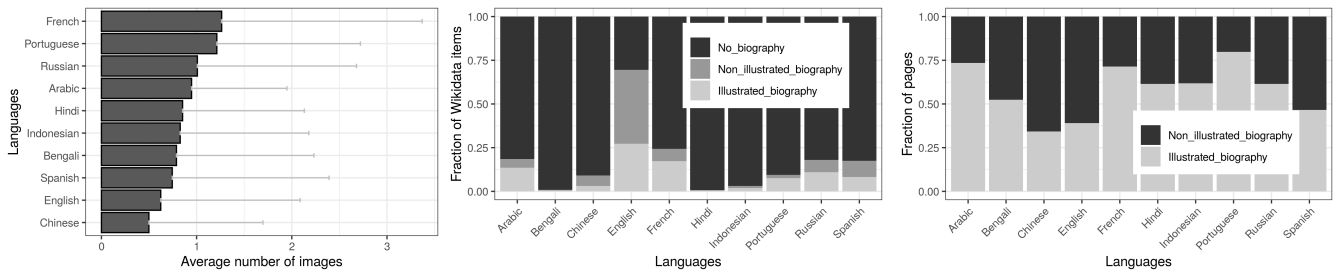


Figure 3: (Left) Distribution of images per page across languages. (Middle) Fraction of different kinds of pages in 6.22 Million items. (Right) Fraction of biographies with images per language.



Figure 4: Left: Low-quality images, i.e., pictures randomly drawn from the image repository of Wikimedia Commons. Right: High-quality images, i.e., pictures tagged with “Quality images” category from Wikimedia Commons.

data for training, and the rest for validation. The classifier achieves 85% accuracy on a balanced test set. The resulting image quality classifier, given any image, outputs a *quality score* in the range  $[0, 1]$ . Figure 4 displays examples of images categorized by this classifier as low or high quality.

### 3.4 Metrics

To get a full, multimodal picture of the gender gap on Wikipedia, across visual and textual content, we define four metrics, reflecting content selection, building, and positioning. Each metric includes two indicators: one capturing the visual aspect of a gap ( $ctype=i$ ) and one capturing the textual or content aspect of the same gap ( $ctype=t$ ). All metrics follow the same generic formula that captures equality between female and male biographies:

$$M_{ctype} = \frac{q(female)}{q(male)} \quad (1)$$

where  $q$  is a generic score/quantity used for analysis (e.g., number of articles),  $q(male)$  refers to that quantity calculated for male biographies, and  $q(female)$  refers to the same quantity for biographies about women.

While Wikidata allows to identify people with 36 possible genders<sup>14</sup>, we decided to focus on comparing males and females (99.9% of biographies with gender attributed). Unfortunately, only 1,223 biographies (0.01%) do not belong to those categories, which are further separated into multiple

identities (e.g., intersex, transgender male, transgender female, agender, hermaphroditism, etc.). This high dissection does not allow us to perform, for those genders, the granular analysis we intend here (observing trends in different languages and types of occupations). We consider this to be a limitation of the current work.

Each metric offers a different perspective on the gender gap, by giving different interpretations of  $q$ . Such formulation allows to express, with one single number (a ratio), the extent of the gender gap as the prominence of the generally dominant category (male), across different dimensions. The closer  $M$  is to 1, the smaller the gender gap, because  $q$  takes similar values for male and female biographies.  $M$  values close to 0 reflect the prominence of male content.  $M$  values above 1 reflect the prominence of female content.

Following Beytía and Wagner (2022), we posit that gender-related *visual* asymmetries can be identified across three stages of visibility production. The collaborative use of images in biographies inevitably produces a *selection* of a set of “illustrated biographies” in each language version. When Wikipedia editors use images to collaboratively characterize each biography (and by extension its gender), they are participating in the content *building* phase. And that process inevitably produces a group of articles which, at the multilingual level, are differentiated by a degree of *visual positioning* of the biographies in various languages.

#### Content Selection

*Monolingual Coverage (MC)*: This metric offers a high-

<sup>14</sup>[https://www.wikidata.org/wiki/Property\\_talk:P21](https://www.wikidata.org/wiki/Property_talk:P21)

level view of the gender bias in Wikipedia articles. Here,  $q$  corresponds to the amount of (illustrated) articles for different genders. More specifically, the  $MC_t$  metric for text reflects the ratio between counts of *all* female biographies vs male biographies, while the  $MC_i$  metric for visual content takes into account *illustrated* articles only:

$$MC_t = \frac{\text{article\_count}(female)}{\text{article\_count}(male)} \quad (2)$$

$$MC_i = \frac{\text{illustrated\_article\_count}(female)}{\text{illustrated\_article\_count}(male)} \quad (3)$$

### Content Building

**Content Quantity (CQ):** This metric reflects the amount of in-article content (text and images) for pages about different genders in Wikipedia. For this metric,  $q$  is defined as the average article length for the textual dimension, and the average volume of images in an article for the visual dimension which we describe in Section 3.

$CQ_t$  is defined therefore as the ratio between the average number of characters in Wikipedia articles about women and the average length of men biographies, while  $CQ_i$  is the ratio between the average number of images in Wikipedia articles about women and the average number of images in men biographies.

$$CQ_t = \frac{\text{avg\_number\_of\_characters}(female)}{\text{avg\_number\_of\_characters}(male)} \quad (4)$$

$$CQ_i = \frac{\text{avg\_number\_of\_images}(female)}{\text{avg\_number\_of\_images}(male)} \quad (5)$$

**Content Quality (CQL):** This metric reflects the quality of article text and images for biographies in Wikipedia. Here,  $q$  is defined as the average quality of an article or its images for different genders.

$CQL_t$  is defined therefore as the ratio between the average article quality for women and the average quality of men biographies, while  $CQL_i$  is the ratio between the average quality of images in Wikipedia articles (refer Section 3) about women and the average quality of images in men biographies.

$$CQL_t = \frac{\text{avg\_article\_quality}(female)}{\text{avg\_article\_quality}(male)} \quad (6)$$

$$CQL_i = \frac{\text{avg\_image\_quality}(female)}{\text{avg\_image\_quality}(male)} \quad (7)$$

### Content Positioning

**Multilingual Coverage (MLC):** This metric provides an overview of the ‘‘popularity’’ of an article (illustrated or unillustrated) across different languages. Here,  $q$  reflects the average number of Wikipedia language editions covering a biography which exists in the language considered.

More specifically, the  $MLC_t$  metric for text reflects the ratio between the average number of languages covering all female biographies divided by the average number of languages covering all the male biographies, while the  $MLC_i$  metric for visual content takes into account the multilingual coverage of *illustrated* articles only:

$$MLC_t = \frac{\text{language\_coverage}(female)}{\text{language\_coverage}(male)} \quad (8)$$

$$MLC_i = \frac{\text{language\_coverage\_illustrated}(female)}{\text{language\_coverage\_illustrated}(male)} \quad (9)$$

## 4 Results

We compute the metrics in Section 3.4 for all articles in our dataset, and report the results in the following subsections.

### 4.1 Quantifying Visual Gaps: Cross-Language Patterns

Figure 5 shows, for the selected languages, the gender bias metrics considered in this study organized according to the three stages of visibility production described in Section 2.

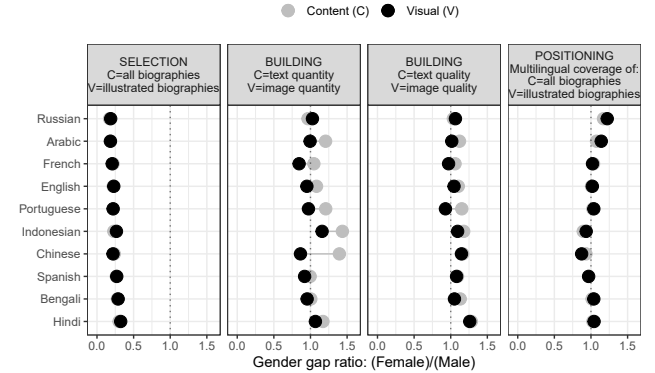


Figure 5: Visual and non-visual gender gaps by Wikipedia language version (in the ten most spoken languages).

Across languages, there is a significant male bias in the *selection* of biographies ( $MC_t$ ), with female biographies covering between 14.7% (Russian) and 22.5% (Hindi) of the articles. Within each language, the selection of illustrated biographies adopts practically the same pattern as that of all biographies. This suggests that the visual bias follows the same pattern as the text bias in the selection stage.

At the content *building* stage, the scenario is quite different. When comparing the amount of text in the articles ( $CQ_t$ ), most languages (except Russian, Spanish and Bengali) have asymmetries in favor of women. This difference is especially noticeable in Indonesian and Chinese. But when we consider the amount of visual content ( $CQ_i$ ), languages tend to show either no asymmetries or a slight male bias (as in French and Chinese). The exception is Indonesian, where women have on average more images than men.

In terms of content quality ( $CQL$ ), we see similar patterns for both text and images. For all languages, the overall quality of women biographies is higher than that of men articles. There exists a clear female bias for image quality only for Chinese, Russian and Bengali (with all other languages having almost the same level of visual quality across genders).

Collectively, these results suggest that, once a biography makes it to Wikipedia, there is a huge dedication from the community to maintain its levels of quality and amount of content, with a stronger focus on text than on images. *As a trend, the textual content was more equitable or women favoring as compared to the visual content.* This speaks to the tedious work of the numerous sub-communities whose mandate, across language editions, is to improve the content and the representation of women in history. By man-



ually inspecting some of the articles with higher content quantity and quality, we found that, for example, most of the articles about women in Russian Wikipedia are curated by the “Project: Feminism”, a community of editors aiming at improving women articles and increasing the number of women biographies rated as Featured articles. This pattern is also present in other languages, which also have projects to improve the coverage and quality of women articles.<sup>15</sup>

Regarding the multilingual *positioning* of content, no significant gender biases are found. However, in some languages (such as Russian and Arabic), women with biographies tend to have higher multilingual coverage than men. In contrast, other languages (such as Chinese and Indonesian) have slightly higher multilingual coverage for men. On these positioning metrics, no major differences can be recognized between the patterns of the content in general ( $MLC_t$ ) and those of the illustrated articles ( $MLC_i$ ).

The multilingual coverage could be understood as a proxy for the notability of the subject of a biography, i.e. how famous the person is across cultures. The greater the number of languages in a biography, the greater its global relevance. Thus, languages with biases in multilingual coverage establish asymmetries with respect to the multicultural relevance of their portrayed men and women.

We interpret these slight asymmetries in an ambivalent way because they could be correlated with a bias in the biographies selection. For example, in Russian and Arabic women average better multilingual coverage than men. To some extent, that could result from the fact that those languages include the lowest proportions of female biographies (14.7% and 15.6%), so they only include biographies of a few very recognized women.

## 4.2 Occupational Patterns in English Wikipedia

To understand these biases more deeply, we will focus on a specific language version (English Wikipedia). This will allow us to decompose the metrics according to the occupations of the people portrayed and to perform a qualitative analysis of the emergence of visual gender biases in specific occupational groups.

Figure 6 shows the gender bias metrics for English Wikipedia. Four trends can be observed. First, most of the male bias comes from the article selection stage. Second, the male bias in the image quantity is the unique male asymmetry after the article selection stage. Third, visual and written content have a clear break down in the content building

<sup>15</sup>In Arabic Wikipedia, The Women’s Wiki Project community works on increasing the quantity and quality of women articles, with a special focus on women from the Arabic world. In Indonesian Wikipedia, the WikiGap project maintains a list of biographies of the great women in history to be improved or translated from other languages. The Hindi Wikipedia community maintains the “WikiProject daughters of Hind” (where Hind stands for the Indian subcontinent), whose aim is to create, maintain or promote articles related to Indian women holding high positions in India and abroad (Figure 1). Beyond WikiProjects, we found that events such as the Wikigap Edithaton organized in 2019 by Bengali Wikipedians, resulted in many articles about women created and improved.

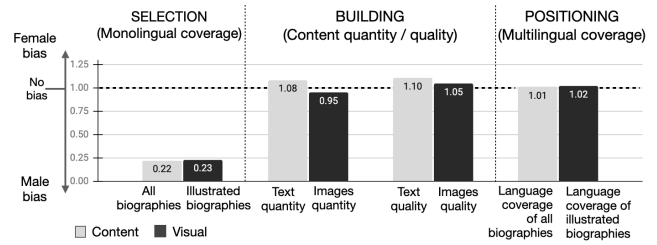


Figure 6: Visual and non-visual gender gaps on English Wikipedia.

metrics: the amount and quality of text favor female biographies,<sup>16</sup> while the quantity of images favors male biographies.<sup>17</sup> Finally, the multilingual coverage of biographies is slightly biased towards women, for all articles (with and without images).<sup>18</sup>

Figure 7 shows the breakdown of these general trends into ten broad categories of occupations. In terms of *selection*, male bias is much more pronounced in professions linked to religion and army than arts and humanities. We can also identify two occupational patterns connected to content *building*. Regarding image quantity, almost all professions show a male bias, although the sports and religion biographies have the opposite trend. With respect to the image quality, most occupations tend to have a slight female bias (especially in law and government), although again biographies of religious figures show a contrasting pattern.

Finally, almost all occupations have similar patterns of visual and non-visual *positioning*. Science and technology biographies stand out because their number of illustrated articles in various languages is substantially more male-biased<sup>19</sup> than their non-visual positioning indicator which is female-biased.<sup>20</sup>

## 4.3 Qualifying Visual Gender Gaps: Exploring Science and Technology Biographies

The above results indicate that visual content patterns are quite independent of written content trends. The science and technology biographies are one example of that. If we only inspect their text, we might think that they do not manifest a relevant gender bias. However, when we include the visual analysis, those same biographies show a male preference associated with the number of images they contain (see Figure 7). We qualitatively review this group of biographies to further explore how a male visual preference in specific occupational groups manifests itself on Wikipedia.

In English Wikipedia, only 14.5% of science and technology biographies are about women. These articles are similar in text length and text quality to those of men. Nevertheless,

<sup>16</sup>We perform Kolmogorov–Smirnov (KS) test to check the significance of the differences between female and male biographies:  $D = 0.06(\text{quantity})$ ;  $D = 0.01(\text{quality})$ ,  $p < 0.0001$  in both.

<sup>17</sup>KS stats:  $D = 0.015$ ,  $p < 0.0001$ .

<sup>18</sup>KS stats:  $D = 0.02(0.03)$ ,  $p < 0.0001$  in both.

<sup>19</sup>KS stats:  $D = 0.1$ ,  $p < 0.0001$

<sup>20</sup>KS stats:  $D = 0.14$ ,  $p < 0.0001$ .

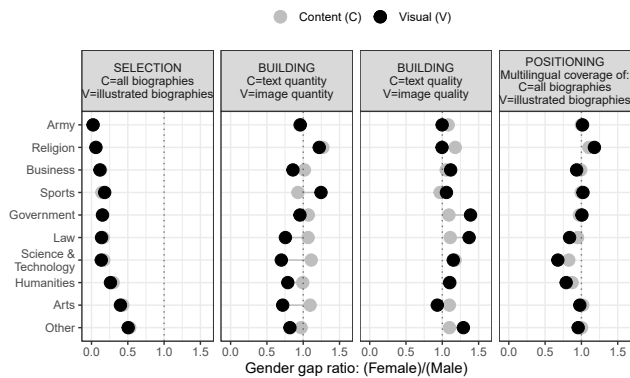


Figure 7: Visual and non-visual gender gaps on English Wikipedia by ten occupational domains.

our data shows that the vast asymmetry in *selecting* articles is combined with a *visual* asymmetry in their building stage: items about men have, on average, more images than those about women. Why is this visual disparity established?

To answer that question, we undertook a qualitative examination of the articles associated with science and technology. Our start-point was a sample of 50,000 randomly selected biographies. We created an interface with this data to find cases of men and women similar according to content indicators<sup>21</sup>. This interface allowed us to segment the biographies by gender and occupational dimension, and also to search for articles according to specific ranges of (1) length, (2) quantity of images, (3) article quality, and (4) multilingual coverage. We focused on identifying cases of men and women who had occupations in science or technology and stand out for their amount of visual content. We then compared the identified cases between genders and analyzed each biography to deepen the qualitative context of the visual differences between men and women.

With this method, we noticed a general trend. The biographies with better visual documentation are mainly about people who made *material productions with high social impact* (e.g., notable inventions or public constructions). For example, Zhang Heng (23 images) invented the seismograph, the water-powered armillary sphere, and improved the water clock. Washington Sheffield (18) introduced toothpaste, and René Laennec (13) developed the stethoscope. Paul Séjourné (14) built long bridges for masonry, while Alan Turing (15) constructed an electromechanical machine called “the bombe.” The visual advantage of these material creations is that they can be directly photographed or drawn, and they also leave a significant trace of objects linked to their production (plans, sketches, formulas, etc.). The latter multiplies the potential for visual documentation in these biographies.

Historically, the production of material objects has been unequally attributed to men, which may partly explain why there is a male visual bias in this broader occupational group.

<sup>21</sup>The interface is available at: [https://public.tableau.com/app/profile/pablo1729/viz/VisualGenderBiasesinWikipedia\\_caseselectionforqualitativeanalysis/Dashboard1](https://public.tableau.com/app/profile/pablo1729/viz/VisualGenderBiasesinWikipedia_caseselectionforqualitativeanalysis/Dashboard1)

On the other hand, we notice that women who stand out for their visual documentation are those that have achieved *significant and continuous professional milestones throughout their lives*, which can be easily represented with photos or symbols associated with their biographical stages.

Grace Hopper, for example, is the woman in our sample with the most images in her biography (10 images). That is because her career as a mathematician and computer scientist was primarily spent in the U.S. Navy, so her biographical milestones are accompanied by images of the symbols representing the naval ranks she obtained (commodore, captain, commander, etc.). Marissa Mayer, in parallel, has good visual documentation because, throughout her life, she has played a leading role in several large information technology companies (such as Google or Yahoo!) and has often appeared in media as a representative of them (see Figure 8). Ornithologist Pamela Rasmussen has been discovering different bird species during her career, and these milestones can be easily documented with explanatory pictures. Finally, the chemist Donna Nelson won academic awards in various ceremonies recorded by media and provided scientific advice for television productions (such as *Breaking Bad*) which facilitated her media exposure.

These women biographies, which stand out for their high visual content on Wikipedia, are not structured as a narrative about *invention and material transformation*, which seems to be the type of story with the greatest visual potential in people linked to science and technology. In their articles, a narrative about *the relevant and publicized milestones of their professional careers* prevails. This form of biographical story undoubtedly has a high *direct* graphic potential, as each milestone can be recorded visually in Wikipedia. Still, it is not usually connected with a significant *indirect* visual trace, such as that left by objects related to material productions and inventions (e.g. plans, sketches, formulas). This relative lack of material traces could partly explain the visual advantage of men in this specific category of articles. An example of these different narrations is shown in Fig. 8.

## 5 Discussion

Our results underscore *the need for analyzing images* to study gender gaps in Wikipedia content. A long tradition of studies focused on advertisements has previously shown that images are critical tools for diffusing gender roles and stereotypes (Goffman 1979; Jones 1991; Kang 1997; Klassen, Jasper, and Schwartz 1993; McLaughlin and Goulet 1999). Our results, complementarily, indicate that visual gender differences on certain dimensions are independent of disparities generated in written content. That suggests that, if the visual analysis is overlooked, cumulative content asymmetries are likely to be less accurately diagnosed and sometimes under- or over-valued.

It is helpful to revisit our results to clarify the above idea. Take the case of arts and sports biographies in English Wikipedia. In both occupational groups, articles about men and women showed, on average, gender equality in the written dimension (considering both the quantity and quality). That could lead us to diagnose the absence of bias and the similarity between both groups in terms of gender patterns.



### René Laennec (1781 - 1826)

Laennec had discovered that the new **stethoscope** was superior to the normally used method of placing the ear over the chest, particularly if the patient was overweight. A stethoscope also avoided the embarrassment of placing the ear against the chest of a woman.<sup>[1]</sup>

Laennec is said to have seen children playing with long, hollow sticks in the days leading up to his innovation.<sup>[2]</sup> The children held their ear to one end of the stick while the opposite end was scratched with a pin, the stick transmitted and amplified the scratch. His skill as a fiddler may also have inspired him. He built his first instrument as a 20 by 2.5 cm hollow wooden cylinder, which he later refined to comprise three detachable parts. The refined design featured a funnel-shaped cavity to augment the sound, separable from the body of the stethoscope.<sup>[3]</sup>

His clinical work allowed him to follow chest patients from bedside to the autopsy table. He was therefore able to correlate sounds captured by his new instrument with specific pathological changes in the chest, in effect pioneering a new non-invasive diagnostic tool. Pulmonary phthisis, for example, was one ailment he could more easily identify using his knowledge of typical and atypical chest sounds.<sup>[4]</sup> Laennec was the first to classify and discuss the terms *rhonchi*, *crackles*, and *egophony* – terms that doctors now use on a daily basis during physical exams and diagnoses.<sup>[5]</sup> Laennec presented his findings and research on the stethoscope to the Academy of Sciences in Paris, and in 1819 he published his masterpiece, *De l'auscultation médiate ou Traité de Diagnostic des Maladies des Pouxins et du Cœur*, in two volumes.<sup>[6]</sup>

Laennec coined the phrase *medicula auscultatoria* (indirect listening), as opposed to the popular practice at the time of directly placing the ear on the chest (*medicula auscultatione*). He named his instrument the **stethoscope**, from the Greek words *στήθος*(*stēthos*) (chest), and *σκοπέω*(*skopēō*) (examination).

The stethoscope quickly gained popularity as *De l'auscultation Médiate* was translated and distributed across France, England, Italy and Germany in the early 1820s.<sup>[7]</sup> However, not all doctors readily embraced the new stethoscope. Although the *New England Journal of Medicine* reported the invention of the stethoscope two years later in 1821, as late as 1885, a professor of medical studies, "to that hath ears to hear, let him use his ears and the Laennec coined the phrase *medicula auscultatoria* (indirect listening), as opposed to the popular practice at the time of directly placing the ear on the chest (*medicula auscultatione*). He named his instrument the **stethoscope**, from the Greek words *στήθος*(*stēthos*) (chest), and *σκοπέω*(*skopēō*) (examination).

Not a stethoscope: "Even the founder of the American Heart Association, L. A. Connor (1866-1900) carried a silk handkerchief with him to place on the wall of the chest for ear auscultation."<sup>[8]</sup>

Laennec often referred to the stethoscope as the "cylinder," and as he neared death only a few years later, he bequeathed his own stethoscope to his nephew, referring to it as "the greatest legacy of my life."<sup>[9]</sup>

The modern **binaural stethoscope** with two earpieces was invented in 1851 by Arthur Levin of Ireland. **Georg Carmin** perfected the design of the instrument for commercial production in 1852, which has become the standard ever since.<sup>[10]</sup>

Other medical contributions [edit]

---

### J. J. Thompson (1856 - 1940)

In May-June 1897, Thomson investigated whether or not the rays could be deflected by an electric field.<sup>[1]</sup> Previous experiments had failed to observe this, but Thomson believed their experiments were flawed because their tubes contained too much gas.

Thomson constructed a Crookes tube with a better vacuum. At the start of the tube was the cathode from which the rays projected. The rays were deflected by a beam by two metal disks – the first of these disks utilized as the anode, the second was connected to the earth. The beam then passed between two parallel aluminum plates, which produced an electric field between them when they were connected to a battery. The end of the tube was a large sphere where the beam would impact on the glass, creating a glowing patch. Thomson painted a scale to the surface of this sphere to measure the deflection of the beam. Any electron beam would collide with some residual gas atoms within the Crookes tube, thereby ionizing them and producing electrons and ions in the tube (space charge), in previous experiments this space charge screened the externally applied electric field. However, in Thomson's Crookes tube the density of residual atoms was so low that the space charge from the electrons and ions was insufficient to electrically screen the externally applied electric field, which permitted Thomson to successfully observe electrical deflection.

When the upper plate was connected to the negative pole of the battery and the lower plate to the positive pole, the glowing patch moved downwards, and when the polarity was reversed, the patch moved upwards.

Measurement of mass-to-charge ratio [edit]

In his classic experiment, Thomson measured the mass-to-charge ratio of the cathode rays by measuring how much they were deflected by a magnetic field and comparing this with the electric deflection. He used the same apparatus as in his previous experiment, but placed the discharge tube between the poles of a large electromagnet. He found that the mass-to-charge ratio was over a thousand times lower than that of a hydrogen ion (H<sup>+</sup>), suggesting either that the particles were very light and/or very highly charged.<sup>[2]</sup> Significantly, the rays from every cathode yielded the same mass-to-charge ratio. This is in contrast to anode rays (now known to arise from positive ions emitted by the anode), where the mass-to-charge ratio varies from anode-to-anode. Thomson himself remained critical of what his work established, in his Nobel Prize acceptance speech referring to "corpuscles" rather than "electrons".

Thomson's calculations can be summarized as follows (in his original notation, using F instead of E for the electric field and H instead of B for the magnetic field):

The electric deflection is given by  $\theta = Fd/mv^2$ , where  $\theta$  is the angular electric deflection, F is applied electric intensity, d is the charge of the cathode ray particles, l is the length of the electric plates, m is the mass of the cathode ray particles and v is the velocity of the cathode ray particles. The magnetic deflection is given by  $\phi = Hcl/mv$ , where

### Sophie Germain (1776 - 1831)

**École Polytechnique** [edit]

In 1784, when Germain was 18, the **École Polytechnique** opened.<sup>[1]</sup> As a woman, Germain was barred from attending, but the new system of education made the lecture notes available to all who asked.<sup>[2]</sup> The new method also required the students to "submit written observations."<sup>[3]</sup> Germain obtained the lecture notes and began sending her work to Joseph Louis Lagrange, a faculty member. She used the name of a former student Monsieur Antoine-Auguste Le Blanc (1710-1789), as she later explained to Gauss, the student attached to a female scientist.<sup>[4]</sup> When Lagrange saw the intelligence of M. Le Blanc, he requested a meeting, and thus Sophie was forced to disclose her true identity. Fortunately, Lagrange did not mind that Germain was a woman,<sup>[5]</sup> and he became her mentor.<sup>[6]</sup>

**Early work in number theory** [edit]

**Correspondence with Legendre** [edit]

Germain first became interested in number theory in 1798 when Adrien-Marie Legendre published *Essai sur la Théorie des nombres*.<sup>[1]</sup> After studying the work, she opened correspondence with him on number theory, and later, *elasticity*. Legendre showed some of Germain's work in the Supplement to his second edition of the *Théorie des nombres*, where he called it *rigoureuse* ("very rigorous"). See also Her work on Fermat's Last Theorem below.<sup>[2]</sup>

**Correspondence with Gauss** [edit]

Germain's interest in number theory was renewed when she read Carl Friedrich Gauss' monumental work, *Disquisitiones Arithmeticae*.<sup>[3]</sup> After three years of working through the exercises and trying her own proofs for some of the theorems,<sup>[4]</sup> she wrote, again under the pseudonym of M. Le Blanc,<sup>[5]</sup> to the author himself, who was one year younger than she.<sup>[6]</sup> The first letter, dated 21 November 1804,<sup>[7]</sup> discussed Gauss' *Disquisitiones* and presented some of Germain's work on Fermat's Last Theorem. In the letter, Germain claimed to have proved the theorem for  $n = p - 1$ , where  $p$  is a prime number of the form  $p = 4k + 3$ .<sup>[8]</sup> However, her proof contained a weak assumption, and Gauss' reply did not comment on Germain's proof.<sup>[9]</sup>

Around 1807 (sources differ),<sup>[10]</sup> during the Napoleonic wars, the French were occupying the German town of Braunschweig, where Gauss lived. Germain, concerned that might affect the fate of Archimedes, wrote to Germain Pernely, a family friend, requesting that he assure Gauss' safety.<sup>[11]</sup> General Pernely sent the sheet of a battalion to meet with Gauss personally to see that he was safe.<sup>[12]</sup> As it turned out, Gauss was fine, but he was confused by the mention of Sophie's name.<sup>[13]</sup>

Three months after the incident, Germain disclosed her true identity to Gauss.<sup>[14]</sup> He replied:<sup>[15]</sup>

How can I describe my astonishment and admiration on seeing my esteemed correspondent M. Le Blanc metamorphosed into this celebrated person... when a woman, because of her sex, our customs and prejudices, encounters infinitely more obstacles than men in familiarising herself with Number Theory's tough problems, yet overcomes these fetters and penetrates that which is most hidden; she doubtless has the noblest courage, extraordinary talent, and superior genius.

Carl Friedrich Gauss [edit]

---

### Marissa Mayer (1975 - )

**Career** [edit]

**Google (1999-2012)** [edit]

After graduating from Stanford, Mayer received 14 job offers,<sup>[1]</sup> including a teaching job at Carnegie Mellon University<sup>[2]</sup> and a consulting job at McKinsey & Company.<sup>[3]</sup> She joined Google in 1999 as employee number 20.<sup>[4]</sup> She started out writing code and overseeing small teams of engineers, developing and designing Google's search offering.<sup>[5]</sup> She became known for her attention to detail, which helped lead her a promotion to product manager<sup>[6]</sup> and later she became director of consumer web products.<sup>[7]</sup> She oversees the layout of Google's web content, unadorned search homepage.<sup>[8]</sup> She was also on the three-person team responsible for Google AdWords, which is an advertising platform that allows businesses to show their product to relevant potential customers based on their search terms. AdWords helped deliver 98% of the company's revenue in the first quarter of 2011.

In 2002, Mayer started the Associate Product Manager (APM) program, a Google mentorship initiative to recruit new talents and cultivate them for leadership roles. Each year, Mayer selected a number of junior employees for the two-year program, where they took on entrepreneurial assignments and mentoring incoming classes.<sup>[9]</sup> Notable graduates of the program include Bill Taylor and Justin Rosenstein.<sup>[10]</sup> In 2005, Mayer became Vice President of Search Products and User Experience.<sup>[11]</sup> Mayer held key roles in Google Search, Google Images, Google News, Google Maps, Google Books, Google Product Search, Google Toolbar, Google, and Gmail.<sup>[12]</sup>

Mayer was the vice president of Google Search Products and User Experience until the end of 2010, when she was asked by then-CEO Eric Schmidt to head the Local, Maps, and Location Services.<sup>[13]</sup> In 2011, she secured Google's acquisition of survey site Zogby for \$125 million. While Mayer was working at Google, she taught introductory computer programming at Stanford and mentored students at the East Palo Alto Charter School.<sup>[14]</sup> She was awarded the Centennial Teaching Award and the Forayby Award from Stanford.<sup>[15]</sup>

**Yahoo! (2012-2017)** [edit]

On July 16, 2012, Mayer was appointed president and CEO of Yahoo!, effective the following day. She was also a member of the company's board of directors.<sup>[1]</sup> At the time of her appointment, Yahoo's numbers had been falling for three years and the company had been through several top management changes. To simplify the bureaucratic process and "make the culture the best version of itself", Mayer launched a new online program called P&U. It collects employee complaints, as well as their votes on problems in the office; if a problem generates at least 50 votes, online management automatically investigates the matter.<sup>[2]</sup> In February 2013, Mayer oversees a major personnel policy change at Yahoo! that required all remote-working employees to convert to in-office roles.<sup>[3]</sup> Having worked from home toward the end of her pregnancy, Mayer returned to work after giving birth to a boy, and built a mother's room next to her office suite—Mayer was consequently criticized for the "recommuting" ban.<sup>[4]</sup> In April 2013, Mayer changed Yahoo's maternity leave policy, lengthening its time allowance and providing a cash bonus to parents.<sup>[5]</sup> CNN noted this was in line with other Silicon Valley companies, such as Facebook and Google.<sup>[6]</sup> Mayer has been criticized for many of her management decisions

enhancements in 2010

Marissa Mayer speaking at the Google "Search On" event in 2010

Marissa Mayer at an interview while working for Google

Michael Angolan and Marissa Mayer at TechCrunch Disrupt

Figure 8: Left: Pictures showing inventions and artefacts on the wiki pages for René Laennec and J. J. Thompson (two men in Science and Technology). Right: Images capturing professional milestones on the biographies of Sophie Germain and Marissa Mayer (two women in Science and Technology).

However, when we analyze the images of these biographies, significant gender asymmetries are found, which also differentiate both occupational groups' diagnoses. Arts biographies display a clear male visual preference (men have biographies with more and better-quality images than women). In contrast, in sports articles, female biographies register on average more and better-quality images. These gender biases establish relevant differences between both groups of biographies. Still, they had not been previously identified due to the dearth of systematic research on images.

Besides, our results illustrate the *complexity of visual asymmetries*. We have documented that gender gaps cannot be reduced to a single bias or content production stage. On the contrary: the visual treatment of gender seems to be a specific "domain" or "register" of communication, which cuts across Wikipedia's main editorial processes. Therefore, to offer a global and systematic perspective on gender bias on this platform, it is preferable to use multiple indicators, associated with many asymmetries along the main stages of content production (selection, building, and positioning). This approach allows us to comprehensively observe how Wikipedia organizes the visibility of each gender, respecting the high level of complexity of this platform and its specific way of organizing visual content (Beytía and Wagner 2022).

The *Glass Ceiling* theory (Wagner et al. 2016) is one of the few approaches that has problematized gender bias in a complex and multidimensional way. For that the con-

ceptual framework, women have higher entry barriers into Wikipedia's record, which causes only very notable women to be documented. That "creaming" phenomenon (Lacireno-Paquet et al. 2002) –where only the most select group is incorporated into the biographical record– would explain that women tend to have, on average, equal or better content and positioning than men. This theory proposes that gender bias in Wikipedia is fundamentally developed at the article selection stage; once women overcome that entry barrier, they obtain similar or better content and positioning than men.

Our results support aspects of this thesis since they show that the male bias is produced mainly in selecting articles. In none of the ten languages analyzed do female biographies represent more than 23% of the record. However, our data also show that there exist consistent male biases after the selection stage. In several languages (French, Chinese, Spanish, Bengali, English, and Portuguese) the image quantity is larger for men. And this is not the only bias after articles selection. Empirical studies have found that men biographies tend to have greater centrality in the hyperlink network between articles (Graells-Garrido, Lalmas, and Menczer 2015; Wagner et al. 2015).

Furthermore, our results dispute that women portrayed on Wikipedia are, on average, equally or more notable than men. In Chinese, Indonesian and Spanish, for example, men's biographies average higher notability (i.e., multilingual coverage) than those of women. Also in some occu-

pational areas (such as science and technology in English Wikipedia), the men portrayed have greater notability than the women. That suggests that, despite the limited inclusion of women's biographies, in not all languages nor occupations women have higher barriers to entry than men.

Taken together, these findings motivate further specification and refinement of the Glass Ceiling theory. Most of the gender bias on Wikipedia indeed comes from the article selection processes (across the ten languages reviewed). However, choosing only very notable and famous women: (1) does not eliminate the existence of systematic male biases after selection (generated in content building and positioning processes), and (2) does not necessarily result in female biographies being equally or more notable than male biographies (in terms of multilingual coverage).

Furthermore, it is necessary to advance *explanations* for these biases. What could be underlying these content patterns? We can suggest a general hypothesis with the help of previous research. It is quite possible that gender inequality in the composition of editors is the main driver of gender biases in content. This is because Wikipedia editors tend to develop an “ingroup bias”, i.e. “to present their own group (the ingroup) in a systematically more favorable way” (Oeberst et al. 2020). For example, editors in a specific language tend to select more information about events, people or things linked to territories that speak the same language (Overell and R ger 2011). And the same is true with respect to gender: research has documented that Wikipedia editors tend to write proportionally more about people and topics related to their own gender identity (Lam et al. 2011; Hinnosaar 2019). Such self-reference to one's own gender could be generating deep biases in content selection, as several studies have shown that within Wikipedia only between 8.5% and 16.1% of editors are women (Glott, Schmidt, and Ghosh 2010; Wikimedia 2011; 2012; Hill and Shaw 2013; Minguill n et al. 2021).

The above argument allows explaining global trends in gender gaps in content. According to it, when there are biases that favor men, this can be partly explained by the fact that the editors of those articles are mostly men. However, that is not enough if we want to explain *ambivalent content*, i.e., when in the same editing environment (e.g., Indonesian Wikipedia) some indicators favor men and others favor women. These cases are precisely the ones that this research has helped to identify, and which generate entirely new questions. Why in Chinese Wikipedia, women tend to have more written content and men more visual content? Why in the sports biographies of English Wikipedia, it is women who tend to have more images and men more written content? These questions are still unanswered, since the literature that evaluates gender bias in a complex way (in multiple dimensions) and considering the written and visual domain has been virtually nonexistent. We think that they should be answered with studies focused on a specific editing environment (e.g., editors of religious biographies on Wikipedia in Spanish) and hopefully including qualitative methods (such as interviews, participant observation, or ethnography) to further understand how that editing environment contributes to producing gender gaps.

In addition to its theoretical contributions, this study has some *practical implications*. As mentioned, many initiatives have been created to balance information between men and women (Women in Red, Women in Green, Wiki Loves Women, WikiWomen's Collaborative, Editatona, Art+Feminism, 500 Women Scientists, etc.). The research in this paper provides granular insights on how such initiatives can balance visual gender gaps: we show the significant types of bias, in which languages they are most strongly expressed, and in which specific groups of occupations. The current study is one of the largest and most comprehensive studies of gender bias on Wikipedia, and the findings of this article (and the accompanying dataset) could guide those efforts, define priorities, and improve equity in Wikipedia content.

For example, our results indicate that gender bias in article selection occurs more strongly in some languages (such as Russian and Arabic). In contrast, asymmetry in the number of images is more present in others (such as French and Chinese). These details in the diagnosis of each language make it possible to guide editorial initiatives more effectively and define which aspects should be prioritized (e.g., incorporating images or the increase in visual quality). Furthermore, our distinction of occupations—which is available for the ten languages in our database—allows these initiatives to focus their efforts further. For example, in English Wikipedia, the editing of female biographies linked to arts could be a priority since they have an unfavorable visual treatment regarding quantity and quality of images.

It is important to list some *technical limitations* of this study. First, we did not consider biographies in all languages, but only in the ten with the most speakers. Second, we did not classify all occupations, but only the 300 most mentioned. Third, we did not conduct an in-depth visual analysis, considering detailed representations in the images. For example, the studies on advertisements we have mentioned usually evaluate six aspects of the pictures that construct female stereotypes: relative size, feminine touch, function ranking, the family, ritualization of subordination, and licensed withdrawal (Goffman 1979; Kang 1997). As a macro approach to shed light on Wikipedia's visual biases, our research does not reach that level of analytical detail. However, future research work could be designed to overcome this limitation.

Finally, our study has some *explanatory limitations*. Our results show relevant differences in the visual content of the articles but do not explain all these differences. On one hand, it would be pertinent to investigate the editors' motive: visual differences could be established, for example, because they (a) decide not to publish pictures, (b) do not have permission to publish them, or (c) do not have access to an image of what they want to illustrate. On the other hand, one could ask whether the biographies' occupational composition explains some of the differences in the broad occupational categories. For example, we know that, in English Wikipedia, women have on average better-quality images than males in sports occupations. Perhaps this is explained by the fact that certain specific sports tend to generate very good quality images, and a higher proportion of women par-

ticipate in those sports. For instance, the case of badminton players on English Wikipedia: they have in general very high quality of images, and the percentage of women dedicated to this sport is higher than the percentage of men. We believe that future research could investigate these explanatory aspects in more specific editing environments or in more focused categories of articles.

## 6 Conclusions

This study's main objective was to advance the development of a general and systematic perspective on visual gender bias in Wikipedia biographies. Previous research has had multiple limitations. They have typically focused on a minimal range of occupations, explored the topic on a monolingual level, and have not compared visual asymmetries with other content gaps. Our study has sought to overcome these limitations: (1) it analyzed visual gender gaps in a vast spectrum of 300 occupations, (2) it studied monolingual and multilingual asymmetries, in English Wikipedia and the ten most spoken languages, (3) it looked at quantitative and qualitative aspects of the images used in those articles, and (4) it compared visual gender asymmetries with non-visual indicators of content, especially metrics of written content.

Our results highlighted some trends across the ten most spoken languages. (1) Much of the male bias develops at the selection stage of biographical articles. (2) Regarding the *amount of content*, these languages tend to have a female bias in the text quantity and a male bias in the image quantity. (3) Our metrics also indicate that in the ten languages female biographies have better *content quality*, and in almost all of them (except French and Portuguese) they average higher *image quality*.

We further explored the case of English Wikipedia and arrived at some findings. (1) The male bias in article selection is significantly more marked in occupations linked to army and religion than in arts and humanities. (2) After the article selection stage, image quantity is the only metric with typical male asymmetry. (3) Visual and written content have a clear break down in the content building metrics: the amount and quality of text favor female biographies, while the image quantity favors male. (4) The multilingual article coverage is biased slightly towards women. Our qualitative analysis of science and technology biographies suggests that the narrative of image-rich articles is different for men and women. Male biographies often concentrate on the invention of objects, which has a greater potential for visual documentation, whereas female articles focus on professional milestones. This narrative analysis, associated with specific occupations and languages, can provide relevant clues to understand the generation of gender biases in Wikipedia biographies. Taken together, these results shed light on the types of visual biases that arise in collaborative Wikipedia editing and provide guidelines for management of these content asymmetries.

## Acknowledgements

We would like to thank Isaac Johnson and Aiko Chou from Wikimedia Foundation for helping with the development of the article and image quality models. Vivek Singh was sup-

ported in part by the US National Science Foundation under Grant SES-1915790. Pablo Beytía was sponsored by a scholarship from the German Academic Exchange Service (DAAD) and the Chilean National Agency for Research and Development (ANID).

## Image Attributions

**Figure 1:** *Rani of jhansi.jpg*, By Lakshmibai, Rani of Jhansi - Gallery, Public Domain, via Wikimedia Commons

**Figure 4:** *Jimmy-Lifton-Mark-Ruffalo-1994.jpg*, By Chadwick Berk, CC BY-SA 4.0, via Wikimedia Commons

*Zhou Zhongyu.jpg*, By Hatsuo Yamagata, Public Domain, via Wikimedia Commons

*Martina Sorbara at Festival of Friends 2013.jpg*, By Tabercil, CC BY-SA 3.0, via Wikimedia Commons

*Lewis Hamilton 2008 Britain 6.jpg*, By Carlina Xavier, CC BY-SA 2.0, via Wikimedia Commons

**Figure 8:** *Rene-Theophile-Hyacinthe Laennec Drawings stethoscope 1819.jpg*, By René-Théophile-Hyacinthe Laennec, Public Domain, via Wikimedia Commons

*Stethoscope-2.png*, By this version: Sonarpulse. original:Huji, Public Domain, via Wikimedia Commons

*Laennecs stethoscope, c 1820. (9660576833).jpg*, By Science Museum London, CC BY-SA 2.0, via Wikimedia Commons

*Porton antigua sede École Polytechnique Sainte Geneviève.JPG*, By Juan Antonio Cordero, CC BY-SA 3.0, via Wikimedia Commons

*Carl Friedrich Gauss.jpg*, By Gottlieb Biermann - Gauß-Gesellschaft Göttingen e.V. (Foto: A. Wittmann)., Public Domain, via Wikimedia Commons

*JJThomsonGasDischargeTubeElectronCavendishLab2013-08-29-17-11-41.jpg*, By Rolf Kickuth CC BY-SA 4.0, via Wikimedia Commons

*JJ Thomson Cathode Ray 2.png*, By J.J. Thomson - Philosophical Magazine, 44, 293 (1897), Public Domain, via Wikimedia Commons

*Thomson cathode ray exp.gif*, By Kurzon , Public Domain, via Wikimedia Commons

*JJ Thomson exp3.gif*, By Shade154 commonswiki, Public Domain, via Wikimedia Commons

*Marissa Mayer, Search On 2010.jpg*, By Robert Scoble, CC BY 2.0, via Wikimedia Commons

*Michael Arrington and Marissa Mayer at TechCrunch Disrupt.jpg*, By Kevin Krejci, CC BY 2.0, via Wikimedia Commons

*Marissa Mayer interview in 2011 III.jpg*, By Magnus Höij, CC BY 2.0, via Wikimedia Commons

## References

- Adams, J.; Brückner, H.; and Naslund, C. 2019. Who counts as a notable sociologist on wikipedia? gender, race, and the “professor test”. *Socius*.
- Agarwal, P.; Redi, M.; Sastry, N.; Wood, E.; and Blick, A. 2020. Wikipedia and westminster: Quality and dynamics of wikipedia pages about uk politicians. In *ACM Hypertext and Social Media*.

- Beytía, P., and Schobin, J. 2020. Networked pantheon: a relational database of globally famous people: Social and behavioural sciences. *Research Data Journal for the Humanities and Social Sciences*.
- Beytía, P., and Wagner, C. 2022. Visibility layers: a framework for systematising the gender gap in wikipedia content. *Internet Policy Review*.
- Beytía, P. 2020. The positioning matters: Estimating geographical bias in the multilingual record of biographies on wikipedia. In *Proceedings of the Web Conference*.
- Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In *IEEE*.
- Erickson, K.; Perez, F. R.; and Perez, J. R. 2018. What is the commons worth? estimating the value of wikimedia imagery by observing downstream use. In *International Symposium on Open Collaboration*.
- Glott, R.; Schmidt, P.; and Ghosh, R. 2010. Wikipedia survey—overview of results. *United Nations University: Collaborative Creativity Group*.
- Goffman, E. 1979. *Gender advertisements*. Macmillan International Higher Education.
- Graells-Garrido, E.; Lalmas, M.; and Menczer, F. 2015. First women, second sex: Gender bias in wikipedia. In *ACM Hypertext & Social Media*.
- Graham, M.; Hogan, B.; Straumann, R. K.; and Medhat, A. 2014. Uneven geographies of user-generated information: Patterns of increasing informational poverty. *Annals of the Association of American Geographers*.
- Gruwell, L. 2015. Wikipedia’s politics of exclusion: Gender, epistemology, and feminist rhetorical (in) action. *Computers and Composition*.
- Hill, B. M., and Shaw, A. 2013. The wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PloS one*.
- Hinnosaar, M. 2019. Gender inequality in new media: Evidence from wikipedia. *Economic Behavior & Organization*.
- Johnson, C., and Williams, B. 2020. Gender and political leadership in a time of covid. *Politics & Gender*.
- Jones, M. 1991. Gender stereotyping in advertisements. *Teaching of Psychology*.
- Kang, M.-E. 1997. The portrayal of women’s images in magazine advertisements: Goffman’s gender analysis revisited. *Sex roles*.
- Klassen, M. L.; Jasper, C. R.; and Schwartz, A. M. 1993. Men and women: Images of their relationships in magazine advertisements. *Journal of Advertising Research*.
- Klein, M.; Gupta, H.; Rai, V.; Konieczny, P.; and Zhu, H. 2016. Monitoring the gender gap with wikidata human gender indicators. In *Symposium on Open Collaboration*.
- Lacireno-Paquet, N.; Holyoke, T. T.; Moser, M.; and Henig, J. R. 2002. Creaming versus cropping: Charter school enrollment practices in response to market incentives. *Educational Evaluation and Policy Analysis*.
- Lam, S. T. K.; Uduwage, A.; Dong, Z.; Sen, S.; Musicant, D. R.; Terveen, L.; and Riedl, J. 2011. Wp: clubhouse? an exploration of wikipedia’s gender imbalance. In *International symposium on Wikis and open collaboration*.
- McLaughlin, T. L., and Goulet, N. 1999. Gender advertisements in magazines aimed at african americans: A comparison to their occurrence in magazines aimed at caucasians. *Sex Roles*.
- Minguillón, J.; Meneses, J.; Aibar, E.; Ferran-Ferrer, N.; and Fàbregues, S. 2021. Exploring the gender gap in the spanish wikipedia: Differences in engagement and editing practices. *Plos one*.
- Oeberst, A.; von der Beck, I.; Matschke, C.; Ihme, T. A.; and Cress, U. 2020. Collectively biased representations of the past: Ingroup bias in wikipedia articles about intergroup conflicts. *British Journal of Social Psychology*.
- Overell, S. E., and Rüger, S. 2011. View of the world according to wikipedia: Are we all little steinbergs? *Journal of Computational Science*.
- Redi, M.; Gerlach, M.; Johnson, I.; Morgan, J.; and Zia, L. 2021. A taxonomy of knowledge gaps for wikimedia projects (second draft). *arXiv:2008.12314*.
- Samoilenko, A.; Lemmerich, F.; Weller, K.; Zens, M.; and Strohmaier, M. 2017. Analysing timelines of national histories across wikipedia editions: A comparative computational approach. In *Proceedings of the AAAI ICWSM*.
- Singh, V. K.; Chayko, M.; Inamdar, R.; and Floegel, D. 2020. Female librarians and male computer programmers? gender bias in occupational images on digital media platforms. *Association for Information Science and Technology*.
- Wagner, C.; Garcia, D.; Jadidi, M.; and Strohmaier, M. 2015. It’s a man’s wikipedia? assessing gender inequality in an online encyclopedia. In *Proceedings of the AAAI ICWSM*.
- Wagner, C.; Graells-Garrido, E.; Garcia, D.; and Menczer, F. 2016. Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ Data Science*.
- Wajcman, J. 2008. Life in the fast lane? towards a sociology of technology and time. *The British journal of sociology*.
- Wikimedia. 2011. Wikipedia editors study: Results from the editor survey. [bit.ly/3umcw2f](http://bit.ly/3umcw2f).
- Wikimedia. 2012. Editor survey: Wikipedia editing experience. [bit.ly/3AwLpmA](http://bit.ly/3AwLpmA).
- Wikimedia. 2020. Wikidata. [wikidata.org/wiki/Wikidata:Main\\_Page](http://wikidata.org/wiki/Wikidata:Main_Page).
- Worku, Z.; Bipat, T.; McDonald, D. W.; and Zachry, M. 2020. Exploring systematic bias through article deletions on wikipedia from a behavioral perspective. In *International Symposium on Open Collaboration*.
- Young, A.; Wigdor, A. D.; and Kane, G. 2016. It’s not what you think: Gender bias in information about fortune 1000 ceos on wikipedia. *Information Systems*.
- Yu, A. Z.; Ronen, S.; Hu, K.; Lu, T.; and Hidalgo, C. A. 2016. Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific data*.
- Zagovora, O.; Flöck, F.; and Wagner, C. 2017. "(weitergeleitet von journalistin)" the gendered presentation of professions on wikipedia. In *ACM Web Science*.