RESEARCH
ARTICLE

OPEN
ACCESS

PEER
REVIEWED

Visibility layers: a framework for systematising the gender gap in Wikipedia content

Pablo Beytía *Humboldt University of Berlin*
Claudia Wagner *RWTH Aachen University*

Abstract: The gender gap in Wikipedia content is a complex phenomenon that comprises several asymmetries, discursive dimensions, and social concerns. However, there is no theoretical framework to organise this complexity consistently. Based on writings by Foucault, Deleuze and Tkacz, we interpret Wikipedia as a 'field of visibility' and provide a framework to systemise its content gaps. Then we use that model to organise the complexity of the content gender gap on Wikipedia, performing a systematic overview of the asymmetries tested in empirical research. We suggest that this analysis is relevant for the effective planning of governance processes that seek to avoid female or non-male subordination in digital platforms' discourses.

Introduction

Wikipedia is currently a central player in the global organisation of knowledge. Not only has it become the fifth most visited website in the world and the most popular one dedicated to building a compendium of knowledge (Alexa Internet, 2019). It did so by disseminating articles in more than 300 languages (Wikipedia, 2020b) and becoming a validated source for scientific research[1]. Within the platform economy, these elements position Wikipedia as the leading infrastructure in compiling knowledge systematically at a global and multicultural level.

Its growing influence has been accompanied by concern about how its encyclopaedic information is being selected, presented, and structured. At the beginning of the project, this worry led mainly to a debate about the *reliability* of its content, where several studies asked whether an encyclopaedia based on open collaboration could achieve the standards of traditional encyclopaedias or other reputable sources of information (Giles, 2005; Bragues, 2007; Clauson et al., 2008; Pender et al., 2008; Rector, 2008; Brown, 2011). Nowadays, however, most studies warn about a different issue: the *systematic asymmetries* of this digital repository. The global claim of these investigations is that Wikipedia tends to disproportionately collect information on specific social groups, most evidently on men, from the Global North –particularly the United States and Western Europe– and who were born in the last century (Nemoto & Gloor, 2011; Overell & Rüger, 2011; Graham et al., 2014; Eom et al., 2015; Gruwell, 2015; Samoilenko et al., 2017; Beytía, 2020).

Among these asymmetries, the tendency to generate more information about men has probably been the most controversial. For some years now, we have known that the composition of Wikipedia *contributors* is hugely unbalanced. Various studies estimated that only between 8.5% and 16.1% of the editors are women (Glott et al., 2010; Wikimedia Foundation, 2011, 2012; Hill & Shaw, 2013; Minguillón et al., 2021) and others pointed out that editorial participation of non-binary gender

---

1. Wikipedia information has been employed to study various topics in social sciences. For example, large demographic trends (Reznik & Shatalov, 2016), associations between languages (Ban et al., 2017; Eom et al., 2015; Ronen et al., 2014), geopolitical instability (Apic et al., 2011), the similitude of interests between countries (Karimi et al., 2015), and migration of famous people during their lives (Menini et al., 2017; Schich et al., 2014).

identities in Wikipedia is virtually non-existent (Stephenson-Goodknight 2017)[2]. Further investigations have suggested that such uneven composition (the so-called 'participation gender gap') is triggering gender biases in articles (a 'content gender gap') since editors tend to write proportionally more about people of and topics related to their own gender identity (Lam et al., 2011; Hinnosaar, 2019).

Therefore, it is not surprising to currently find a substantial gender bias in the number of biographies in Wikipedia. If we consider the 6.22 million articles registered on all its language versions in February 2020, men represented 70.1%, women 19.9%, and 10% had no gender record (Beytía et al., forthcoming). Despite the platform allows up to 36 gender types to be attributed (Wikidata, 2021), non-binary identities only represent about 0.01% of those biographies[3] (Beytía et al., forthcoming).

The above disparity in the number of biographies is the basic fact of the 'content gender gap' in Wikipedia. And it has given rise to significant reactions. Within Wikipedia, numerous initiatives have emerged to reinforce women's content: Women in Red, Women in Green, Wiki Loves Women, WikiGap, WikiWomen's Collaborative, WikiHerStory, Editatona, and several others. A few recent events have also been organised to boost content about people with non-binary gender identity[4]. In addition, non-profit organisations (such as Art+Feminism or 500 Women Scientists) have appeared to counteract the content trends by organising massive editing events. Furthermore, accountability projects such as 'Wikidata Human Gender Indicators' (Klein et al., 2016) have supported this work, providing weekly updated information on gender disparities in Wikipedia content and detailing specific statistics about countries, cultures, and historical periods.

In the meantime, the gender disparity in the number of biographies has led to a deeper matter: is the content gender gap only associated with the selection of articles, or is it also expressed in other discursive features of content (e.g., articles' length, quality, and centrality)? A few studies have tried to systematically answer this question, assessing *multiple aspects* in which the content gender gap could be expressed in Wikipedia (Graells-Garrido et al., 2015; Wagner et al., 2015, 2016).

---

2. This '*participation* gender gap' has been explained by several factors. For example: (1) Wikipedia has a geek identity that is unappealing to certain genders, (2) its editorial openness tends to include difficult people in its communities, which is especially annoying for women, (3) editors dismiss concern for editorial diversity because of the supposed freedom to participate, (4) the project builds on previous infrastructures with high gender exclusion—such as the enlightened encyclopaedia and the open Internet (Reagle 2013; Ford & Wajcman 2017).

3. 1,223 biographies for a total of 6.22 million.

4. For example, the LGBTQ edit-a-thon on 26 February 2021.

And they have been able to show features of discursive disparities that had perhaps been imagined but scarcely explored empirically.

The former investigations have also extended the 'content gender gap' concept and opened new theoretical questions. Before them, this disparity could have been understood very concretely as the *unequal proportion of articles* between two or more genders, which considered gender *representation* as the central concern. Now, there has been a dissection of the gap in many imbalances expressed in various discursive features. This situation has had several consequences. First, it has led to a *more abstract understanding* of the gender gap, now defined as a 'systematic asymmetry' in the way two or more genders are treated and presented (Wagner et al., 2016 p. 2). Second, it has implied linking this gap with *new concerns*: not only the gender 'representation' but also its 'characterisation' and 'structural placement' in the content. Finally, this gap's greater complexity implies starting a theoretical discussion about *the appropriate dimensions and indicators* to understand and measure this disparity; so far, there is no agreement on the elements that compose it[5].

In short, we currently have a more complex approach to the content gender gap, with a more abstract definition and linked to new concerns. But *that approach still lacks a clear structure of analysis*. That is because there is no conceptual framework to integrate all the empirical results on this subject in a theoretically coherent way. For example, there is no explanation of what are the *processes and stages* of knowledge organisation involved in the formation of this gap, what *agents* are implicated in the creation of each gender imbalance, in what sense these content asymmetries are components of *the same phenomenon* (i.e., 'the content gender gap'), or in what way they are *'framing'* communication about gender. In other words, we need a theoretical framework that justifies the joint analysis of the researched content asymmetries, links these disparities to agents and their editing processes in Wikipedia, provides meaning to the combined attention to aspects such as the representation, characterisation and structural placement of each gender.

This article aims to propose such a conceptual model and then use it to systematise the empirical literature's findings, associating each discursive asymmetry with specific editorial processes, agents involved in content development, and modes of

5. Among the few investigations conducted, some analytical dimensions do not match, and others are similar but using different names or indicators. Graells-Garrido et al. (2015) analysed asymmetries in three dimensions: meta-data, language, and network structure; Wagner et al. (2015) opted for four dimensions: coverage, structure, lexicon, and visibility. Wagner et al. (2016) focused on five aspects: notability, topical focus, linguistic asymmetry, structural properties, and meta-data presentation.

framing communication.

The theoretical systematisation of this gap is relevant in the current background of the platform economy, i.e., the social process where activities to exchange, share and collaborate that are associated with production and consumption increasingly rely on digital infrastructures (Algan et al., 2013; Fuster Morell et al., 2020). The main result of this research is a *'multi-layer analysis'* of how one of the oldest and more influential peer production platforms is shaping women's visibility. This inspection is fundamental in several ways. First, because Wikipedia has *broad consequences* on the current distribution of knowledge in a global and multilingual level. Second, it is a *paradigmatic example* of a collaborative platform in this 'new economy,' and could give us insight into the challenges that newer platforms might have in the future. Third, it is a well-studied case, which allows us to recognise the *high degree of complexity* involved in discursive gender inequalities within digital infrastructures. We think that this complexity awareness is a prerequisite to plan *effective processes of platform governance* with gender orientation. Finally, this is an excellent case to reflect on how *digital discourses cunningly spread gender asymmetries*. Wikipedia is a platform powerfully designed to include diverse editors and editorial standpoints, which nevertheless introduces substantial content exclusions. It is necessary to clarify that unexpected result by building a comprehensive diagnosis, which brings together all types of content asymmetries and points out which specific editorial processes and agents are generating these discursive disparities.

In the following section, we introduce the 'Visibility layers' model: the conceptual framework that we propose to analyse content gaps in Wikipedia. Subsequently, we organise, within that model, the empirical literature on gender biases in Wikipedia content. In that section, we systematically review ten types of discursive asymmetries that are classified into three 'stages of visibility production': content selection, building and positioning. After providing a general and multidimensional overview of the content gender gap in Wikipedia, we highlight the main contributions of this theoretical-empirical analysis and discuss some of its implications.

Visibility layers: a framework for systematising content gaps

Suppose someone opens a journal about Wikipedia and finds these three headlines:

1. 70.1% of biographies are about men (Beytía et al., forthcoming).

2. Articles on women focus more on family relationships (Wagner et al.,
    2016).
3. The most central articles (or best connected with other articles) are
    predominantly about men's lives (Graells-Garrido et al., 2015).

All these statements address gender imbalances in Wikipedia content. However,
they do not have much more in common. Each of them emphasises *different as-
pects of the content*, developed by *distinct agents*, and has *distinctive discursive con-
cerns*.

The first statement refers to the process of *selecting articles*, which in Wikipedia in-
volves only those editors who propose articles and the reviewers who accept
them. The underlying concern, in that case, is the low discursive 'representation' of
non-male persons. However, the second assertion refers to the *topics* developed
within the articles, which concerns the specific group of editors who have dedicat-
ed themselves to writing and discussing those articles. Their primary worry would
be the 'characterisation' of women in their biographies. The third statement refers
to the *position* of a group of articles within the whole content system of the ency-
clopaedia in a specific language, potentially involving all the editors (humans and
bots) who have edited Wikipedia in that language. In that case, the main concern
seems to be the 'structural placement' of non-male persons. How could these three
statements be understood, then, as descriptions of the same phenomenon?

However, recent scientific literature considers all these statements associated with
Wikipedia's *content gender gap* (Graells-Garrido et al., 2015; Wagner et al., 2015,
2016). Instead of stating that this is a conceptual mistake, we would like to pro-
pose that specialists have begun to understand this gap as a *complex* discursive
phenomenon –i.e., composed of multiple elements with different characteristics,
which are related to each other in a temporally variable manner (Luhmann,
1999)–. This situation demands a theoretical framework capable of organising that
growing complexity coherently. This section will propose a theoretical model for
systematising the Wikipedia content gender gap.

Theoretical background

To build this model, we will draw mainly on two sources: (1) a theory that under-
stands knowledge organisations as *regimes of visibility*, and that goes back to
Michel Foucault's studies in the 1960s and 1970s (Foucault, 2013, 2012), although
it was later systematised by Gilles Deleuze (1988) and used by Nathaniel Tkacz
(2007) to analyse the reorganisation of power in Wikipedia; and (2) the official in-
formation on how this platform works, which can be found on several websites of

the encyclopaedia (e.g., Wikipedia, 2020f, 2020c, 2020a) and has been systema-
tised in some publications (Lih, 2009; Jemielniak, 2014; Beytía & Müller, 2019).

Our starting point can be two classical studies made by Foucault on social
processes of the 17th and 18th century: *History of Madness* and *Discipline and Pun-
ish* (Foucault, 2012, 2013). In the first one, Foucault analysed the emergence of
health institutions such as the general hospital, the correctional house, and the
asylum. In the second one, he focused on the birth of the prison as a modern insti-
tution of surveillance.

Following Deleuze (1988), there is a kind of parallelism in both investigations.
Foucault understands both the general hospital and the prison as 'architectures',
but not because they are aggregations of stones or means of confinement, but in-
stead because they are *ways of organising light*. His position is consistent with Aris-
totle's perspective, which distinguished architecture from other material arts due
to its focus on *form* rather than matter (Aristotle, 2005). Following Foucault and
Deleuze's interpretation of his work, modern hospitals and prisons could be de-
fined as 'light sculptures' or 'fields of visibility' since they establish a distribution
of light and shadow over a specific area. They are places that 'make you see' cer-
tain content and articulate a collective, multi-sensory perception of a topic. The
prison would be the field of visibility of the crime (the crime brought to light),
while the general hospital would be the field of visibility of the madness (the way
it is brought to light). Thus, both institutions can be understood as *complexes of
multisensorial conditions* (not only optical but also auditory, tactile, and other forms
of perception) that make feasible distribution of visibility towards specific objects
or contents.

In 2007, Nathaniel Tkacz published a small essay where he applied this theory to
the scrutiny of Wikipedia. Based on Foucault and Deleuze's abstractions –which
highlight the *formal* (rather than material) organisation of places such as hospitals
and prisons–, he proposed that Wikipedia can also be interpreted as an 'architec-
ture' or 'field of visibility'. The characteristics of that architecture are what would
distinguish it from previous encyclopaedias. For example, Wikipedia includes new
tools that *make visible* editorial processes that were previously hidden: there is free
access to the complete evolution of each article, and each edited topic includes a
discussion forum (talk page) where divergent positions are expressed and recorded
(Tkacz, 2007).

For our purpose, the central point is that Tkacz associated Wikipedia with the the-
ory of visibility put forward by Foucault and Deleuze. On that basis, we could un-

derstand this encyclopaedia as a specific way of illuminating a type of content. However, the general hospital was shedding and structuring light on the madness, while the prison was shedding and structuring light on the crime. What would be the specific matter illuminated by Wikipedia? Since this platform defines itself as 'a written compendium of knowledge' (Wikipedia, 2020d), we could say that it specifically sheds and structures light on knowledge. Then, *Wikipedia would be a field of visibility on knowledge.*

Though, what issues would concern a theory about visibility in Wikipedia? From the end of Tkacz's essay, we can distinguish four elements: (1) *the form of illumination*: how things are enlightened and hidden; (2) *the infrastructure arrangements*: 'how what we can see relates to a politics of arrangement, of architecture and design'; (3) *the social relations*: 'how what is made visible is bound up in relations of power'; and (4) *the framing*: 'how understandings of knowledge are shaped by the architecture that enables them' (Tkacz, 2007, p. 17).

It follows from the above that, in order to properly understand Wikipedia's visibility, it would be necessary to have a conceptual framework that (1) serves to distinguish what is multi-sensory visible from what is not and to identify degrees of visibility in the content; (2) considers how the infrastructure of this platform (involving software, design, editing flow, and other aspects) is associated with processes to manage this visibility; (3) associates the visibility processes with the agents involved, to whom their power relations could be monitored; and (4) specifies to what extent these 'architectural' processes are generating 'frames' in Wikipedia articles –i.e., ways to select aspects of the perceived reality and make them more salient in the communication (Entman, 1993, 2007).

The visibility layers model

In Figure 1, we propose a model that attempts to meet all the above requirements and could potentially be used to analyse *any* content gap in Wikipedia (not only the content gender gap but also geographical, historical, occupational, or any other kind of information asymmetry)[6]. This model's basic idea is that *what lies behind the content gaps is a repartition of multi-sensory visibility, which, in a complex information system like Wikipedia, is developed by overlapping several editorial layers.* Each layer is identified with specific *editing processes* made by a particular group of *agents involved* (left part of the diagram). Additionally, each one is part of a *stage of visibility production* associated with a specific *mode of framing communication* (right

6. This model is developed with more details in Beytía (forthcoming).

part of the diagram). Most of these layers are expressed directly in the content perceived by Wikipedia users (these are *'manifest layers'* of content). In contrast, others refer to observable processes—visible and with discourse-generating potential—which are not displayed in the official encyclopaedic content (these are *'latent layers'*, such as those generated in the processes of article suggestion and editorial discussion on talk pages). We suggest that *the set of all these overlapping layers develops greater or lesser visibility to content about specific issues or social groups in Wikipedia.*
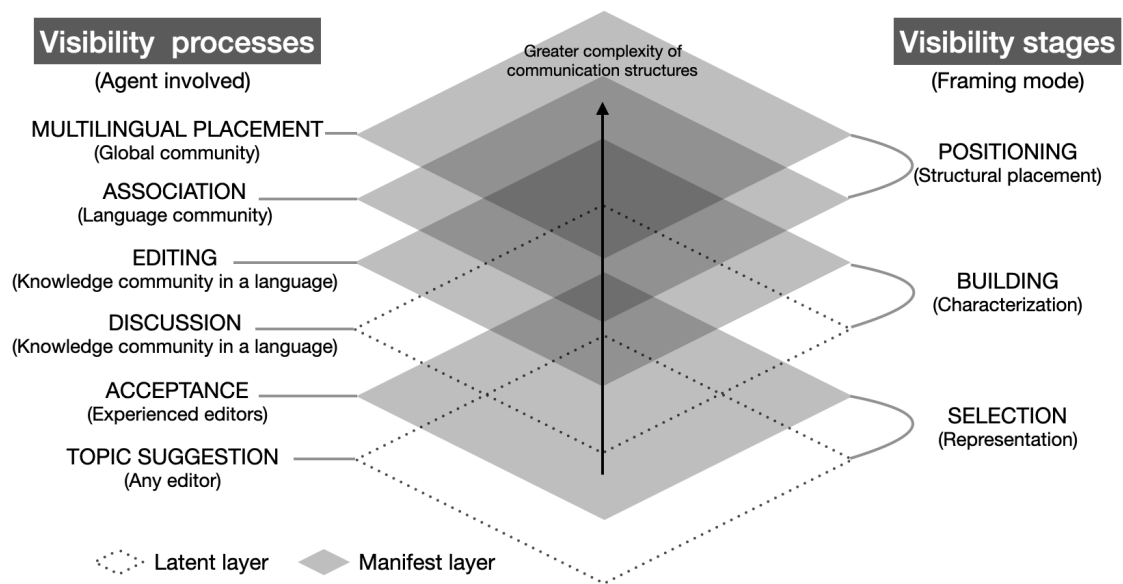

FIGURE 1: Visibility layers: a model for analysing content asymmetries in Wikipedia.

Each of these layers is also at different levels of communicative complexity. At the base of the content elaboration is the simplest communicative process, the *topic suggestion*, where any editor (even those who are not registered in Wikipedia) can write the outline of an article and propose it for the compendium. The suggestions generated in this process only form a 'manifest' layer of content once they receive an encyclopaedia experienced editor's *acceptance*. Both processes –topic suggestion and acceptance– are then responsible for the content *selection* stage, where what is at stake, in terms of framing, is the *representation* of a particular topic –either a person, an animal, a thing or an event– or a set of topics within the encyclopaedia –e.g., notable women–.

Once a topic has been selected, the process of collaborative *editing* of the article begins. That is done by those editors interested in the topic, who wish to collaborate, and are part of Wikipedia's building in a specific language. This communicative activity no longer involves only an editor and an the contributing reviewer(s)

but could include tens, hundreds, or even thousands of editors. When these editors need to discuss editorial changes or resolve disagreements, they can interact in a 'talk page'—a parallel forum designed exclusively for the editorial *discussion* of a specific article. The content *building* stage combines these two linked processes, editing and discussion, in dynamics that are responsible for selecting the written and audio-visual information to be placed within each article and choosing how that information is structured and presented. In terms of framing, what is at stake at this stage is the *characterisation* of a specific topic or set of topics—e.g., the collective portrait of notable women.

When looking at Wikipedia as a system of well or bad connected articles, which are also often available in several languages, two more visibility layers emerge. Among the whole community of editors in a specific language, there is a process of *association* between articles, which involves, on the one hand, the 'classification' of content into thematic categories, and on the other hand, a more spontaneous 'linking' of content produced by connecting articles through hyperlinks. Furthermore, each topic can be developed in one or more languages (Wikipedia is available in more than 300 idioms), which significantly changes its multicultural influence capacity. Therefore, we also distinguish a layer of visibility management called *multilingual placement*, managed by the global Wikipedia community in different languages. Both association and multilingual placement processes can significantly amplify the inequality of content about social groups (Beytía, 2020; Beytía & Schobin, 2020). They are part of the content *positioning* stage, which in terms of framing, puts the *structural placement* that a topic or set of topics has into play. In this stage, information classification schemes are established—for example, by grouping articles into encyclopaedic categories. In addition, the processes of multilingual dissemination and content association establish patterns of centrality and periphery. Articles that are more multilingual in scope and receive more references from other articles constitute a discursive centre. In contrast, those that are elaborated in few languages and receive few references establish a discursive periphery.

Visibility on Wikipedia is built from all these *overlapping* layers of content organisation. Admittedly, article selection is a crucial process to define what is and what is not illuminated. However, it is only the beginning of a series of processes that make each article or collection of articles more or less visible. Our analytical framework seeks to highlight these diverse processes that generate *nuances of illumination*. They are found when one also observes the articles' building and the discursive macro-phenomena of positioning. Following this idea, it is still relevant to wonder 'how many articles are there?' or 'in what proportion?', but new ques-

tions also arise. For example: 'what is the content of these articles?', 'what topics are they associated with?', 'what is their level of centrality?', or 'how is their coverage across different language versions of Wikipedia?'.

Organising the content gender gap: ten asymmetries in three visibility stages

In recent years, multidimensional studies on the content gender gap have multiplied questions about gender visibility in the same direction we have just pointed out. Most of them have specifically compared content about men and women since those groups of articles comprise 99.9% of the biographies with gender reported in Wikipedia (Beytía et al., forthcoming) and thus allow for analyses with greater granularity and depth. Instead of simply asking 'what is the proportion of women's biographies?', these studies have expanded the interrogations to include numerous discursive phenomena, such as the way language is expressed, the visual materials that are selected, the topics covered in the articles, the quality of information, the classification of content, and the network structure of the articles. In all these discursive aspects, the existence of gender asymmetries has been properly tested.

This section aims to systematically document how these gender asymmetries can be coherently integrated within the theoretical model previously proposed. Our main question is: in which layer of visibility should we classify the content disparities already tested in the empirical literature? The value of this classification, as we have suggested, is that it allows us to link each information imbalance to a group of editing agents, a stage of visibility production, and a specific way of frame communication, and then see all the asymmetries as participants in a unitary process of organising visibility.

We will divide the main literature findings into the three already explained stages of visibility production: content selection, building, and positioning.

Content selection

At this stage, mainly two types of asymmetries have been studied. In terms of framing, both are associated with the degree of *representation* of each gender within the totality of the content on remarkable human beings.

The '*article coverage*' asymmetry consists of an uneven proportion of articles about people representing two or more genders (although it has generally been estimated only by comparing women and men). It is commonly reported that women have a modest record of articles in Wikipedia, covering approximately 13.2% to 22.5% of

biographies (Graells-Garrido et al., 2015; Yu et al., 2016; Beytía et al., forthcoming). Some studies have found relevant differences between languages. For example, in Russian Wikipedia, 14.4% of the biographies are about women, while in Hindi 22.5% (Beytía et al., forthcoming). Other investigations suggest that this imbalance decreases if one looks at the biographies present in many languages (Wagner et al., 2015; 2016).

This imbalance has been analysed in two main ways. Most research simply reports a high disproportion in the selection of articles. However, other studies suggest that it is necessary to compare Wikipedia's level of female representation with that of other sources or actual populations to know if there is a male bias *specifically attributable to Wikipedia*. They reported that Wikipedia has a slightly higher proportion of women biographies than some large biographical databases (Freebase, Human Accomplishments, and Pantheon), although articles on women are more likely to be missing than the *Encyclopaedia Britannica* (Reagle & Rhue, 2011; Wagner et al., 2015). The gender distribution of biographies has also been compared with the gender distribution of actual social groups. For example, one study compared US sociologists with their Wikipedia record and found that men are twice as likely as women to have a biographical article (Adams et al., 2019), confirming the coverage asymmetry's relevance.

The second disparity tested at this visibility stage is the '*deletion*' asymmetry, understood as a disproportionate elimination of articles on one or more genders. At Wikipedia, content selection is not a definitive process. Editors maintain the possibility to propose the deletion of articles, which initiates an internal process of debate (called 'Article for Deletion' discussion) that can lead to four outcomes: the article is kept, deleted, merged with another, or redirected to another (Taraborelli & Ciampaglia, 2010).

Studies that have tested gender trends in this process have found no systematic asymmetries: there appear to be no further deletions of content about women or more nominations to remove content about women (Adams et al., 2019; Worku et al., 2020). That suggests that the main challenge for female representation is in the initial selection process and not in this subsequent review stage.

Content building

Primarily five types of asymmetries have been studied at this stage, which in terms of framing are linked to the *characterisation* of the genders in the encyclopaedic content.

The first is the '*writing length'* asymmetry and refers to the extent of texts about each gender, usually measured by the number of words or characters that all articles about people with that identity have on average. Empirical evaluations of this aspect do not suggest that there is a preference for the masculine. Instead, they point out that women's articles tend to be longer than articles about men (Graells-Garrido et al., 2015; Wagner et al., 2015). However, they also warn that this phenomenon could be a side effect of the article coverage asymmetry since Wikipedia only records articles about very notable women but may cover fewer notable men (Wagner et al., 2015).

Second, the '*lexical*' asymmetry can be defined as the unequal association of each gender with specific words or categories of words repeated in the texts. For example, it has been found that the most typical words in men's articles are associated with sports. In contrast, the most familiar words in women articles are more varied and related to gender, achievements, and family (Graells-Garrido et al. 2015). Other approaches note that more abstract (or non-explicit) terms are used to describe the positive aspects of men and the negative aspects of women (Wagner et al. 2016), which could be interpreted as a tendency to value men better through implicit language.

The '*topical*' asymmetry, understood as the unequal association of each gender with typical issues addressed in the articles' text, has also been studied. In this regard, it has been found that women's biographies tend to focus more on gender, social relationships, and family characteristics than men's biographies (Wagner et al., 2016).

The '*visual*' asymmetry is the fourth disparity related to content building. It can be defined as an imbalanced use of images when comparing different genders' content. Research has found that articles on men and women do not differ significantly in the percentage of pages with images (Beytía et al., forthcoming). However, in the ten most widely spoken languages, men's biographies tend to have more images, and female biographies average better visual quality. Additionally, some occupations such as art, humanities, science and technology tend to have a better visual record for men (Beytía et al., forthcoming). Furthermore, visual asymmetry could be displayed in non-biographical content. For example, research on German Wikipedia found that images used to describe occupations have an evident gender asymmetry: almost half of the images from the profession articles show men, and only around 12% depict women (Zagovora et al., 2017).

Finally, the existence of a '*source*' asymmetry, conceived as an unequal use of refer-

ences in the construction of content about people of different genders, has been examined. The relevance of this aspect lies in the idea that the use of appropriate sources is a significant factor in the quality of encyclopaedic articles (Nielsen, 2007; Lewoniewski et al., 2017). One study compared article sources among male and female CEOs, finding that women's biographies have more references and more diverse sources (Young et al., 2016).

Content positioning

At this stage, we can distinguish three asymmetries examined in the literature, which are associated with the *structural placement* of content in terms of framing.

First, we call '*classification*' asymmetry to the systematic association of each gender with thematic categorisation, metadata patterns, and other forms of information classification. So far, we know that men outstand (even more) in sports categories and women in arts categories (Graells-Garrido et al., 2015). Additionally, the selection of notable women in Wikipedia –as opposed to the selection of men– seems to be correlated with the fact that they are married to someone also notable (Graells-Garrido et al., 2015). The latter could be a sign that female notability is sometimes subordinated to male notability processes.

Second, we find a '*network position'* asymmetry, defined as gender imbalances concerning the position of biographies in the network of hyperlinks between articles. This asymmetry is usually estimated using network centrality coefficients (in-degree, k-coreness, PageRank, or other measures) and assortativity indicators (which calculate the preference of articles for linking to similar ones, in this case in terms of gender). Studies have shown that men tend to be more central in the hyperlinks network –at least in Wikipedia in English, Russian, and German– and there is assortativity and (pro men) asymmetry of connectivity across genders (Wagner et al., 2015). Furthermore, biographies with the highest centrality are predominantly about men, and this asymmetry is stronger in Wikipedia than that obtained from simulations of networks with similar structural characteristics (Graells-Garrido et al., 2015).

Finally, the existence of a '*multilingual notability'* asymmetry has been evaluated, that is, the presence of a gender disproportion in the degree of dissemination of biographies in multiple languages. That asymmetry is usually measured by calculating, for each gender, the average number of language editions of Wikipedia in which their biographies have been published. One study found that women are, on average, slightly more notable than men in English Wikipedia, even controlling for

occupation and year of birth (Wagner et al., 2016). That could be explained by the fact that only very prominent women are included in Wikipedia, while men have fewer access barriers, and therefore average a lower level of notability in different languages (Wagner et al., 2016). However, this trend of Wikipedia in English could not be generalised to other languages. A recent study calculated the number of languages in which all biographies included in all Wikipedia editions are available and estimated a reverse scenario: biographies about men are on average in 1.87 languages and those about women in 1.46 (Beytía et al., forthcoming). Additionally, that study estimated that (very scarce) biographies about non-binary genders are on average in 4.19 Wikipedia editions.

General overview

Table 1 organises the ten types of content asymmetries reviewed in this section and links them to the three stages of visibility proposed in our conceptual framework. This procedure allows us to clarify the processes, editorial agents, and framing modes that would be involved in the construction of each asymmetry, and also to understand how these content disparities are participating in an overall process of organising gender visibility.

TABLE 1: Visibility stages and gender asymmetries in Wikipedia

VISIBILITY STAGE	VISIBILITY PROCESS(ES)	AGENT(S) INVOLVED	ASYMMETRY	FRAMING MODE
Selection	Topic suggestion / Acceptance	Any editor / Reviewers (experienced editors)	Article coverage	Representation
			Deletion*	
Building	Editing	Knowledge community in a language	Writing length*	Characterisation
			Lexical	
			Topical	
			Visual	
			Source*	
Positioning	Association	Language community	Classification	Structural placement
			Network position	
	Multilingual placement	Global community	Multilingual notability	

* = These asymmetries have not shown a systematic preference for male content in the empirical tests.

This organisation of asymmetries is the result of a systematic review of the empirical literature on the content gender gap in Wikipedia. However, it should not be

understood as a finished or static structure. An advantage of having a theoretical model is that it can be used to classify new empirical findings, which are then quickly associated with processes, stages, agents and modes of framing communication. Therefore, we understand this classification of asymmetries as a starting point from which new discursive gaps found in future research can be located.

Conclusion

We started this article with a precise diagnosis: the content gender gap of Wikipedia is currently understood as a complex phenomenon since it includes multiple asymmetries, measured in different ways, which are sometimes related to each other and are expressed variably through time. That has led to this gap having a more abstract definition –a 'systematic asymmetry' in the way that two or more genders are treated and presented (Wagner et al., 2016, p. 2)– and being linked to more diverse concerns, such as the representation, characterisation, and structural placement of women. However, empirical research has not used the same analytical structure since there is no theoretical framework to organise this new level of complexity coherently. We suggested that an appropriate framework should link each asymmetry to specific editing processes, editorial agents, and modes of framing communication, as well as explain in what sense all these asymmetries could be considered aspects of the same phenomenon (i.e., the content gender gap).

Our proposal (based on previous works by Foucault, Deleuze, and Tkacz) was to understand Wikipedia as a *field of visibility on knowledge*, that is, an 'architecture' or way of organising light that establishes a distribution of multi-sensory visibility on different topics and thus articulates a collective perception of knowledge. That would not be a neutral articulation, but a distribution that 'makes one see' certain information and hides other, distinguishing furthermore nuances of visibility in the topics considered. This visibility theory allowed us to frame the analysis of content gaps since the processes, agents, and framing modes that emerge in organising the light on Wikipedia should be the same that articulate a specific visibility on each over- or under-represented information topic. Therefore, we have suggested using the *Visibility layers* model as a general theoretical framework to analyse content gaps in Wikipedia. Additionally, we used it to organise the current complexity of the content gender gap (which, from this perspective, could be defined as a systematic asymmetry in the way the multi-sensory visibility of two or more genders is organised).

Using this *multi-layered model*, we associated ten gender asymmetries investigated

in the empirical literature with stages of visibility production, editorial processes, participating agents, and modes of framing communication. This multidimensional analysis, which only compared male and female content (due to literature limitations), indicated a *clear male dominance in Wikipedia's selection of articles*. This selection asymmetry has been *reinforced by modes of building and positioning information*. Articles about females, on average, do not have reduced writing length, source usage, or visual quality. However, they tend to characterise women based on ascribed elements –such as gender and family relations–, contain fewer images and have inferior visual coverage in biographies linked to the arts, humanities, science and technology. Articles about women have less dissemination in multiple languages and their multilingual coverage is sometimes associated with the fact that depicted women had a relevant relationship with a notable man. Moreover, female articles tend to have a more peripheral position in the inter-article referencing network.

At the beginning of this article, we stated that this multi-layered perspective of the content gender gap is relevant in the platform economy's context. Wikipedia is probably the most successful platform dedicated to compiling knowledge globally and multiculturally, which has broad cultural consequences in world society. It is also one of the oldest examples of discourse organisation with peer production mechanisms, making it a significant case for understanding the social outcomes of the collaborative platforms that distinguish this 'new economy'. Additionally, its content asymmetries (especially those related to gender) have been studied in-depth, which allow us to more accurately observe the degree of complexity that discursive inequalities are developing in digital platforms. We think it is necessary to face this high level of complexity with theoretical frameworks that enable us to examine it in an organised way. That is essential for planning governance and content moderation processes that can aspire to the effective production of discourses without a female or non-male subordination. For this reason, we consider fundamental that cyberfeminism theorisation—i.e., the systematic effort to expose, criticise and explain the different relationships of female subordination in the digital society (Jackson & Jones, 1998; Reverter-Bañón, 2013; Oksala, 2017)—and other practices that aspire to avoid the undervaluing of gender identities on the internet, adopt a multi-layered perspective for the analysis of content asymmetries. That would be helpful to expand and connect their expositions, critiques, and explanations of the discourses that are subordinating gender identities in the digital society.

Acknowledgments

References

Adams, J., Brückner, H., & Naslund, C. (2019). Who Counts as a Notable Sociologist on Wikipedia? Gender, Race, and the "Professor Test. *Socius*, *5*. https://doi.org/10.1177%2F2378023118823946

Alexa Internet. (2019). *Alexa Top 500 Global Sites*. Amazon. https://www.alexa.com/topsites

Algan, Y., Benkler, Y., Fuster Morell, M., & Hergueux, Jj. (2013). Cooperation in a Peer Production Economy Experimental Evidence from Wikipedia. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2843518

Apic, G., Betts, M. J., & Russell, R. B. (2011). Content Disputes in Wikipedia Reflect Geopolitical Instability. *PLoS ONE*, *6*(6), e20902. https://doi.org/10.1371/journal.pone.0020902

Aristotle, & Coughlin, G. (2005). *Physics, or, Natural hearing*. St. Augustine's Press.

Ban, K., Perc, M., & Levnajić, Z. (2017). Robust clustering of languages across Wikipedia growth. *Royal Society Open Science*, *4*(10), 171217. https://doi.org/10.1098/rsos.171217

Beytía, P. (2020). The Positioning Matters: Estimating Geographical Bias in the Multilingual Record of Biographies on Wikipedia. *Companion Proceedings of the Web Conference 2020*, 806–810. https://doi.org/10.1145/3366424.3383569

Beytía, P. (forthcoming). *A Digital Setting of Human History. Social Memory and Discursive Power in the Biographical Storage of Wikipedia* [Doctoral thesis]. Humboldt University of Berlin.

Beytía, P., Agarwal, P., Redi, M., & Singh, V. K. (2021). *Visual Gender Biases in Wikipedia: A Systematic Evaluation across the Ten Most Spoken Languages* [Preprint]. SocArXiv. https://doi.org/10.31235/osf.io/59rey

Beytía, P., & Müller, H.-P. (2019). *Towards a Digital Reflexive Sociology: Exploring the Most Globally Disseminated Sociologists on Multilingual Wikipedia* [Preprint]. SocArXiv. https://doi.org/10.31235/osf.io/3pfrv

Beytía, P., & 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*, *5*(1), 50–65. https://doi.org/10.1163/24523666-00501002

Bragues, G. (2007). Wiki-Philosophizing in a Marketplace of Ideas: Evaluating Wikipedia's Entries on Seven Great Minds. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.978177

Brown, A. R. (2011). Wikipedia as a Data Source for Political Scientists: Accuracy and Completeness

of Coverage. *PS: Political Science & Politics*, *44*(02), 339–343. https://doi.org/10.1017/S1049096511
000199

Clauson, K. A., Polen, H. H., Boulos, M. N. K., & Dzenowagis, J. H. (2008). Scope, Completeness, and
Accuracy of Drug Information in Wikipedia. *Annals of Pharmacotherapy*, *42*(12), 1814–1821. https://d
oi.org/10.1345/aph.1L474

Deleuze, G., & Hand, S. (1988). *Foucault*. University of Minnesota Press.

Entman, R. M. (1993). Framing: Toward Clarification of a Fractured Paradigm. *Journal of
Communication*, *43*(4), 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Entman, R. M. (2007). Framing Bias: Media in the Distribution of Power. *Journal of Communication*,
57(1), 163–173. https://doi.org/10.1111/j.1460-2466.2006.00336.x

Eom, Y.-H., Aragón, P., Laniado, D., Kaltenbrunner, A., Vigna, S., & Shepelyansky, D. L. (2015).
Interactions of Cultures and Top People of Wikipedia from Ranking of 24 Language Editions. *PLOS
ONE*, *10*(3), e0114825. https://doi.org/10.1371/journal.pone.0114825

Ford, H., & Wajcman, J. (2017). 'Anyone can edit', not everyone does: Wikipedia's infrastructure and
the gender gap. *Social Studies of Science*, *47*(4), 511–527. https://doi.org/10.1177/03063127176921
72

Foucault, M. (2012). *Discipline and punish: The birth of the prison* (2nd Vintage Books ed). Vintage
Books.

Foucault, M., & Khalfa, J. (2013). *History of madness*. Routledge.

Fuster Morell, M., Espelt, R., & Renau Cano, M. (2020). Sustainable Platform Economy: Connections
with the Sustainable Development Goals. *Sustainability*, *12*(18), 7640. https://doi.org/10.3390/su12
187640

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, *438*(7070), 900–901. https://doi.or
g/10.1038/438900a

Glott, R., Schmidt, P., & Ghosh, R. (2010). *Wikipedia survey–overview of results* (Vol. 8). United
Nations University: Collaborative Creativity Group.

Graells-Garrido, E., Lalmas, M., & Menczer, F. (2015). First Women, Second Sex: Gender Bias in
Wikipedia. *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15*, 165–174. ht
tps://doi.org/10.1145/2700171.2791036

Graham, M., Hogan, B., Straumann, R. K., & Medhat, A. (2014). Uneven Geographies of User-
Generated Information: Patterns of Increasing Informational Poverty. *Annals of the Association of
American Geographers*, *104*(4), 746–764. https://doi.org/10.1080/00045608.2014.910087

Gruwell, L. (2015). Wikipedia's Politics of Exclusion: Gender, Epistemology, and Feminist Rhetorical
(In)action. *Computers and Composition*, *37*, 117–131. https://doi.org/10.1016/j.compcom.2015.06.00
9

Hill, B. M., & Shaw, A. (2013). The Wikipedia Gender Gap Revisited: Characterizing Survey Response
Bias with Propensity Score Estimation. *PLoS ONE*, *8*(6), e65782. https://doi.org/10.1371/journal.pon
e.0065782

Hinnosaar, M. (2019). Gender inequality in new media: Evidence from Wikipedia. *Journal of Economic
Behavior & Organization*, *163*, 262–276. https://doi.org/10.1016/j.jebo.2019.04.020

Jackson, S., & Jones, J. (Eds.). (1998). *Contemporary feminist theories*. New York University Press.

Jemielniak, D. (2014). *Common knowledge? An ethnography of Wikipedia*. Stanford University Press.

Karimi, F., Bohlin, L., Samoilenko, A., Rosvall, M., & Lancichinetti, A. (2015). Mapping bilateral information interests using the activity of Wikipedia editors. *Palgrave Communications*, *1*(1), 15041. https://doi.org/10.1057/palcomms.2015.41

Klein, M., Gupta, H., Rai, V., Konieczny, P., & Zhu, H. (2016). Monitoring the Gender Gap with Wikidata Human Gender Indicators. *Proceedings of the 12th International Symposium on Open Collaboration*, 1–9. https://doi.org/10.1145/2957792.2957798

Lam, S. (Tony) K., Uduwage, A., Dong, Z., Sen, S., Musicant, D. R., Terveen, L., & Riedl, J. (2011). WP:clubhouse?: An exploration of Wikipedia's gender imbalance. *Proceedings of the 7th International Symposium on Wikis and Open Collaboration - WikiSym '11*, 1. https://doi.org/10.1145/2038558.2038560

Lewoniewski, W., Węcel, K., & Abramowicz, W. (2017). Analysis of References Across Wikipedia Languages. In R. Damaševičius & V. Mikašytė (Eds.), *Information and Software Technologies* (Vol. 756, pp. 561–573). Springer International Publishing. https://doi.org/10.1007/978-3-319-67642-5_47

Lih, A. (2009). *The Wikipedia revolution: How a bunch of nobodies created the world's greatest encyclopedia* (1st ed). Hyperion.

Luhmann, N. (1999). *Die Gesellschaft der Gesellschaft*. Suhrkamp Verlag.

Menini, S., Sprugnoli, R., Moretti, G., Bignotti, E., Tonelli, S., & Lepri, B. (2017). Ramble On: Tracing Movements of Popular Historical Figures. *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 77–80. https://aclanthology.org/E17-3020

Minguillón, J., Meneses, J., Aibar, E., Ferran-Ferrer, N., & Fàbregues, S. (2021). Exploring the gender gap in the Spanish Wikipedia: Differences in engagement and editing practices. *PLOS ONE*, *16*(2), e0246702. https://doi.org/10.1371/journal.pone.0246702

Nemoto, K., & Gloor, P. A. (2011). Analyzing Cultural Differences in Collaborative Innovation Networks by Analyzing Editing Behavior in Different-Language Wikipedias. *Procedia - Social and Behavioral Sciences*, *26*, 180–190. https://doi.org/10.1016/j.sbspro.2011.10.574

Nielsen, F. A. (2007). Scientific citations in Wikipedia. *ArXiv:0705.2106 [Cs]*. http://arxiv.org/abs/0705.2106

Oksala, J. (2017). Feminism and Power. In A. Garry, S. J. Khader, & A. Stone (Eds.), *The Routledge companion to feminist philosophy*. Routledge, Taylor & Francis Group.

Overell, S. E., & Rüger, S. (2011). View of the world according to Wikipedia: Are we all little Steinbergs? *Journal of Computational Science*, *2*(3), 193–197. https://doi.org/10.1016/j.jocs.2011.05.006

Pender, M. P., Lasserre, K., Kruesi, L., Del Mar, C., & Anuradha, S. (2008). *Putting Wikipedia to the test: A case study*. https://espace.library.uq.edu.au/data/UQ_193433/SLA_Paper.pdf?Expires=1645363234&Key-Pair-Id=APKAJKNBJ4MJBJNC6NLQ&Signature=eTlChDiZJj44ckpbcEKUkONumChouRTx1h1g5Xb~H7s1FCDJwfYEwj9A5Pgq-QSe-ih~trRoMnUtrNkHG3lnxtXLhgnjWey5b6-CjG3Qdwg~ZqkFMvPuDqY65AvKRlKY9gk5ZdMcd-gijcQg2TgyDKE7C~sITFCxsPvvP-Q7qqY9iFBR7yzl6NPlFovnqzCFCFvjerKJK2Ok5xy3ohB3Rv7NI9SBIcmDUFpRuqWjyqXZ~98cN~aaMqpnYIP~N5WljookisxpxrZ-OjaZrjpQZOnq-AlwiCZcF32-orD-TDEPs98bqKh11XKp9y5jD4J69h1KqUj~mxpCIaAPjp9lIg__

Reagle, J. (2013). "Free as in sexist?" Free culture and the gender gap. *First Monday*, *18*(1). https://do i.org/201301131820

Reagle, J., & Rhue, L. (2011). Gender bias in Wikipedia and Britannica. *International Journal of Communication*, *5*. https://ijoc.org/index.php/ijoc/article/view/777/631

Rector, L. H. (2008). Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference Services Review*, *36*(1), 7–22. https://doi.org/10.1108/0090732 0810851998

Reverter-Bañón, S. (2013). *Ciberfeminismo: De virtual a político. 10*(2), 451–461.

Reznik, I., & Shatalov, V. (2016). Hidden revolution of human priorities: An analysis of biographical data from Wikipedia. *Journal of Informetrics*, *10*(1), 124–131. https://doi.org/10.1016/j.joi.2015.12.00 2

Ronen, S., Gonçalves, B., Hu, K. Z., Vespignani, A., Pinker, S., & Hidalgo, C. A. (2014). Links that speak: The global language network and its association with global fame. *Proceedings of the National Academy of Sciences*, *111*(52), E5616–E5622. https://doi.org/10.1073/pnas.1410931111

Samoilenko, A., Lemmerich, F., Weller, K., Zens, M., & Strohmaier, M. (2017). Analysing Timelines of National Histories across Wikipedia Editions: A Comparative Computational Approach. *ArXiv:1705.08816 [Cs]*. http://arxiv.org/abs/1705.08816

Schich, M., Song, C., Ahn, Y.-Y., Mirsky, A., Martino, M., Barabási, A.-L., & Helbing, D. (2014). A network framework of cultural history. *Science*, *345*(6196), 558–562. https://doi.org/10.1126/scienc e.1240064

Stephenson-Goodknight, R. (2017). *Gender diversity mapping project – Diversity Conference 2017*. Wikimedia. https://www.youtube.com/watch?v=GgTkIE9UGsk

Taraborelli, D., & Ciampaglia, G. L. (2010). Beyond Notability. Collective Deliberation on Content Inclusion in Wikipedia. *2010 Fourth IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshop*, 122–125. https://doi.org/10.1109/SASOW.2010.26

Tkacz, N. (2007). Power, Visibility, Wikipedia. *Southern Review: Communication, Politics & Culture*, *40*(2), 5–19.

Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *ArXiv:1501.06307 [Cs]*. http://arxiv.org/abs/1501.06307

Wagner, C., Graells-Garrido, E., Garcia, D., & Menczer, F. (2016). Women through the glass ceiling: Gender asymmetries in Wikipedia. *EPJ Data Science*, *5*(1), 5. https://doi.org/10.1140/epjds/s13688-0 16-0066-4

Wikidata. (2021). *Property talk:P21—Wikidata*. Wikidata. https://www.wikidata.org/wiki/Property_tal k:P21

Wikimedia Foundation. (2011). *Wikipedia Editors Study: Results from the Editor Survey*. Wikimedia Foundation.

Wikimedia Foundation. (2012). *Editor Survey 2012: Wikipedia editing experience*. Wikimedia Foundation. https://upload.wikimedia.org/wikipedia/commons/8/81/Editor_Survey_2012_-_Wikipedi a_editing_experience.pdf

Wikipedia. (2020a). *Category:Wikipedia policies and guidelines*. Wikipedia. https://en.wikipedia.org/w/

index.php?title=Category:Wikipedia_policies_and_guidelines&oldid=965863417

Wikipedia. (2020b). *List of Wikipedias*. Wikipedia. https://en.wikipedia.org/w/index.php?title=List_of_Wikipedias&oldid=973389715

Wikipedia. (2020c). *Wikipedia:Five pillars*. Wikipedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:Five_pillars&oldid=864696430

Wikipedia. (2020d). *Wikipedia:Wikipedia is an encyclopedia*. Wikipedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:Wikipedia_is_an_encyclopedia&oldid=859157990

Worku, Z., Bipat, T., McDonald, D. W., & Zachry, M. (2020). Exploring Systematic Bias through Article Deletions on Wikipedia from a Behavioral Perspective. *Proceedings of the 16th International Symposium on Open Collaboration*, 1–22. https://doi.org/10.1145/3412569.3412573

Young, A., Wigdor, A., & Kane, G. (2016, December). It's Not What You Think: Gender Bias in Information about Fortune 1000 CEOs on Wikipedia. *ICIS 2016 Proceedings*. https://aisel.aisnet.org/icis2016/SocialMedia/Presentations/15

Yu, A. Z., Ronen, S., Hu, K., Lu, T., & Hidalgo, C. A. (2016). Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific Data*, *3*(1), 150075. https://doi.org/10.1038/sdata.2015.75

Zagovora, O., Flöck, F., & Wagner, C. (2017). "(Weitergeleitet von Journalistin)": The Gendered Presentation of Professions on Wikipedia. *Proceedings of the 2017 ACM on Web Science Conference*, 83–92. https://doi.org/10.1145/3091478.3091488