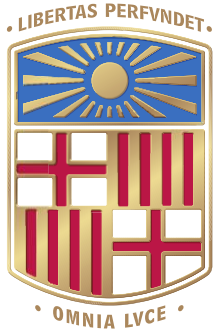
WikiLibCon25
Wikimedia+Libraries International Convention 2025
15 - 17 January 2025
El Colegio de México (COLMEX), Mexico City, Mexico

The Effects of Outlier Data (About Gender and Intersectionalities) in Wikidata on Wikipedia's Main Page: Results of the 'Cover Women' Project

Núria Ferran-Ferrer, Miquel Centelles, Laura Fernández, Marina Salse, Enric Senabre, Julià Minguillón y Francisco Kuggler + TEAM of CHATERS AND FEMINIST USER GROUPS



UNIVERSITAT DE BARCELONA

Motivation of our research 1: Wikipedia

- 7th most visited website in the world.
- Wikipedia is the world's largest free knowledge platform.
- Over 300 language editions and millions of daily visitors.
- The most popular page on Wikipedia is its main page:
  - 2024 4,4 bilion visitors.
  - The front page serves as the entry point, showcasing articles that represent the encyclopedia.
  - Importance of the front page in shaping public perception.

Motivation of our research 2: Gender Gap

- **Acknowledged issue**: recognized by researchers and the Wikimedia Foundation.
  - Bibliographic review (2007-2022): https://doi.org/10.3145/epi.2023.nov.17

- **Key challenges in content**:
  - **Underrepresentation:** Women's biographies make up less than 25% of entries on the Spanish Wikipedia.
  - **Content deletion:** These biographies face higher rates of rapid deletion or contentious debates.

- **Key Challenges in Participation:**
  - **Low participation:** Female contributors represent only 10%-15% on some Wikipedias. *(No data available for non-binary or other gender identities.)*
  - **Early withdrawal:** Women are more likely to stop editing during the first weeks compared to men.

Out data on Gender Gap: https://www.ub.edu/wikiwomen/2024/07/17/wiki-women/

Framework of our research projects

- **Objective:** Investigate and reduce gender and intersectionality biases in Wikipedia.

- **Theoretical framework**:
  - **LIS discipline:** Assessing **knowledge organization systems (KOS)** including the taxonomy of Wikipedia and the ontologies of Wikidata (Centelles & Ferran-Ferrer, 2024).
  - **Communication theories:**
    - **Gatekeeping Theory**: Highlights how decisions about featured content on Wikipedia are influenced by editors' biases. (Barzilai-Nahon, 2009)
    - **Agenda setting**: issues frequent covered in the main page as media agenda (McCombs&Shaw, 1972)
  - **Feminist media studies** (Haervey, 2020) and **Intersectional Feminism** (Crenshaw,1991): Explores how overlapping systems of oppression affect content visibility: Ethnicity, profession, mother tongue, nationality, religion, etc.

- **Ultimate goal: Advocate for positive changes that promote a more equitable and diverse online information environment.**

- Publications and outcomes of our research projects: https://www.ub.edu/wikiwomen/publications/

Cover Women project

- **Objective**: Investigate and reduce gender and intersectionality biases in Wikipedia's main page.
  - 7 WP editions: English, German, Catalan, French, Italian, Portuguese and Spanish.
  - Daily front pages from 2013-2023.
- **Approach**: Wikipedia maing page (COVER) as a new media.

Characteristic	Traditional media	New media	Wikipedia	Wikipedia main page
Content distribution	Centralized and scheduled.	Descentralized, digital and interactive.	Digital and open for global free access.	Curated and highlights selected content.
Audience engagement	Passive or limited active interaction.	Active engagement throught features like comments and sharing	Users can actively edit and discuss content.	Limited interactivity, users suggest changes.
Information sources	Professional contributiors.	Mix of professional and user-generated content.	Decentralized network of volunteers.	Content selected by a tem of editors.
Regular updates	Scheduled updates at specific intervals.	Continuos, real-time updates.	Continyously updated.	Updated daily to reflect current events.
Mass reach	Broad audience, often regional or global.	Global accessibility via Internet.	Global audience in multiple languages.	Global reach as the central access point.

The Wikipedia Main Page blends characteristics of:

- new media: dynamic updates, global reach.
- traditional media: curation, limited engagement.

Methodology: Cover Women Project

Triangulation with 2 approaches

Qualitative approach

a) Insights from community volunteers:

- **Objective:** Understand editorial decision-making, strategies and bias affecting content on WP front page (*tacit knowldege*).
- **Tasks:** 5 interviews for each 7 editions (35 in-deph interviews). Transcription and content analysis.
- **Currently:** Interviewed 13 Wikipedia editors (ENG, SPN, CAT) to understand editorial decision-making.

b) Newsrooms guidelines:

- **Objective:** How does gatekeeping impact on peer production of knowledge within editorial guidelines (*explicit knowledge*)
- **Tasks:** Identify and content analyse the editorial guidelines shaping content selection.
- **Currently:**17 guidelines ENG and 7 for the SP analysed. Trying to find the guidelines for other editions.

c) Main-page content analysis:

Quantiative approach

- **Objective:** Reveal bias trends on gender and intersectional bias in the content featured on WP front page.
- **Tasks:** Identify biographies in each language edition for 10 years, with OpenRefine, enhancing data with Wikidata properties such as P21 sex or gender, P206 occupation, P172 ethnic group, etc.
- **Currently:** Archiving the past 6 months of front pages from 7 editions.

Analyzed 2013–2023 Spanish and English editions (22,924 biographiesacross 4,218 Wikipedia front pages). Forthcoming WP in German and Portugues (arquivo.pt but not exhaustive).

a)**Key results**: Insights from volunteers

- **Diversity is not sufficiently prioritized in the editorial process:**
  - Inequalities in the representation of global knowledge on the platform is perpetuated.

- **Positive actions or efforts to mitigate bias seen as introsion:**
  - Editorial decisions are heavily influenced by volunteers' personal interests and autonomy.
  - Efforts to reduce bias, like specific templates or guidelines, have limited impact, possibly due to resistance from experienced editors who dominate decision-making and favor the status quo.

a)Specific results1: Insights volunteers

- **Recognition of diversity:**
  - Volunteers are perceived to have a skewed interest in certain topics and figures that align with their personal preferences, limiting the diversity of content.

  - This exacerbates the underrepresentation of women, minority ethnic groups, unconventional professions, and less prominent regions such as Africa and Latin America.

- **Infrastructure and community culture**
  - More experienced editors with higher edit counts hold greater influence over content selection.
  - This perpetuates a culture where dominant voices maintain the status quo, discouraging new and diverse contributors.

a)Specific results2: Insights volunteers

Participation bias and content bias have an impact on diversity:

PARTICIPATION

If the volunteer body is homogenous, it will edit on specific topics close to its interests.

CONTENT

If the contents presented are not diverse, this could discourage new readers or editors, as they won't find space for topics of their interest.

DIVERSITY

less diversity among participants leads to less diversity in content, which further discourages diverse participation.

UNIVERSITAT DE BARCELONA

WIKIMEDIA FOUNDATION

a)Specific results 3: Insights volunteers

- **Editorial guidelines and Content selection:**
  - Although guidelines exist to minimize bias, their effectiveness depends heavily on the topics that editors choose to focus on. For instance, the prioritization of politicians and writers over other professions demonstrates a lack of active effort to ensure diverse representation.

  - In the Spanish Wikipedia, some guidelines explicitly recommend prioritizing women, lesser-known countries, unusual professions, or exceptional biographies when equally valid options exist, but the impact of these policies appears limited.

b)Key results: Newsroom guidelines

- *Newsroom: central place to gather news to be publish.*
- *Editorial board: to decide what to be said.*

- Both editions prioritize quality and neutrality criteria,
  - but the Spanish edition explicitly aims to address gender and representational biases.

- The English version has more developed and technical guidelines,
  - which may restrict access to novice editors.

b)Specific results ENG: Newsroom guidelines

- **Detailed and specific guidelines:**
  - There are 17 general and specific guides associated with major sections like "Today's Featured Article," "Did You Know?", "In the News," etc.
  - Criteria for "Featured Articles" include professional writing quality, factual accuracy, neutrality, and inclusion of verifiable references.
  - Article quality is prioritized over significance, which may result in representational biases.
- **Selection Process:**
  - Administrators have the authority to make last-minute changes to the front page.
  - Selections are based on the contributions and experience of volunteer editors.
- **Efforts to mitigate bias:**
  - Templates exist to minimize bias effects, but their effectiveness depends on the contributions of editors.
  - Thematic diversity is often limited by the personal interests of the most active editors.
- **Language and Technical Accessibility:** Extensive use of abbreviations and complex coding can hinder accessibility for new editors.

b)Specific results SP: Newsroom guidelines

- **Fewer but Similar Guidelines:**
  - There are 7vguides associated with sections such as "Artículo Destacado" (Featured Article), "Efemérides" (On This Day), and "Actualidad" (In the News).
  - Quality criteria include neutrality, verifiable references, stability, and the inclusion of multimedia content.
- **Focus on Diversity:**
  - Guidelines for the "Efemérides" section prioritize the representation of women, lesser-known countries, unusual professions, and exceptional events when similar options are available.
  - This reflects a more explicit effort to address representational biases compared to the English edition.
- **Selection Based on Editing Experience:**
  - A minimum number of contributions (50–500 depending on the task) is required to participate in the selection of featured or good articles.
- **Lower Technical Complexity:**
  - The use of abbreviations and complex coding is more limited, making guidelines more accessible.

c)**Key results**: front page content

- **Gender Representation**:
  - ENG WP: 29% women, 0.2% other genders.
  - Spanish WP: 18% women, 0.2% other genders.

- **Intersectional Bias**:
  - Dominance of Western figures and male-associated professions (e.g., politicians, writers).
  - Underrepresentation of marginalized ethnicities, particularly from Africa and Latin America.
- **Editorial Practices**:
  - Volunteer editors' biases often reflect societal norms.
  - Limited systemic efforts to promote diverse representation.

c)Specific results1: front page content

- **Occupation (P106)**:
  - English Wikipedia:
    - Noticeable bias in favor of politicians;
    - «Politician» (appears 4,835 times, surpassing the second category, «Writer», almost doubling its frequency).
  - Spanish Wikipedia:
    - «Politician» category also occupies the first position, with 172 appearances;
    - «Writer» ranks second with 126 appearances.
    - This mirrors the disposition of the English edition.

Wikimedia Commons: 36th G8, 2010 . Photo by: Pete Souza

c)Specific results2: Cover Women project

Religion (P140):

- English Wikipedia:
  - 15 of 20 main values associated with Abrahamic religions (Judaism, Christianity, Islam).
  - Predominance of Christianity, which contrasts sharply with the world distribution of the main religious groups, where Christianity has 30.1% of the believers and Islam 25.1% (Wikipedia.org, 2023).

- Spanish Wikipedia:
  - Similar pattern, 15 of 20 main values associated with Abrahamic religions.
  - Altogether, the values linked to Christianity represent 67.62% of the data set.

Wikimedia Commons: Star with Symbols of religions

UNIVERSITAT DE BARCELONA

WIKIMEDIA FOUNDATION

c)Specific results3: front page content

Country of origin and citizenship (P495 and P27):

- English Wikipedia:
  - The United States of America represents a predominant 49.06%.
  - 86.15% of the total in the following 10 values: United Kingdom, Japan, France, Italy, India, Dutch East Indies, Spain, Canada, Germany and United States.
  - Africa and Latin America are notably underrepresented, with only 7 of the 87 countries in these two regions being mentioned in this attribute.

- Spanish Wikipedia:
  - In the case of "Country of origin", the USA maintains its predominance (52.39%).
  - 14 current European countries correspond to 20.99% of the values.
  - Spanish-speaking countries or regions represent only 13.55% of the total values, with Spain contributing half of this representation.

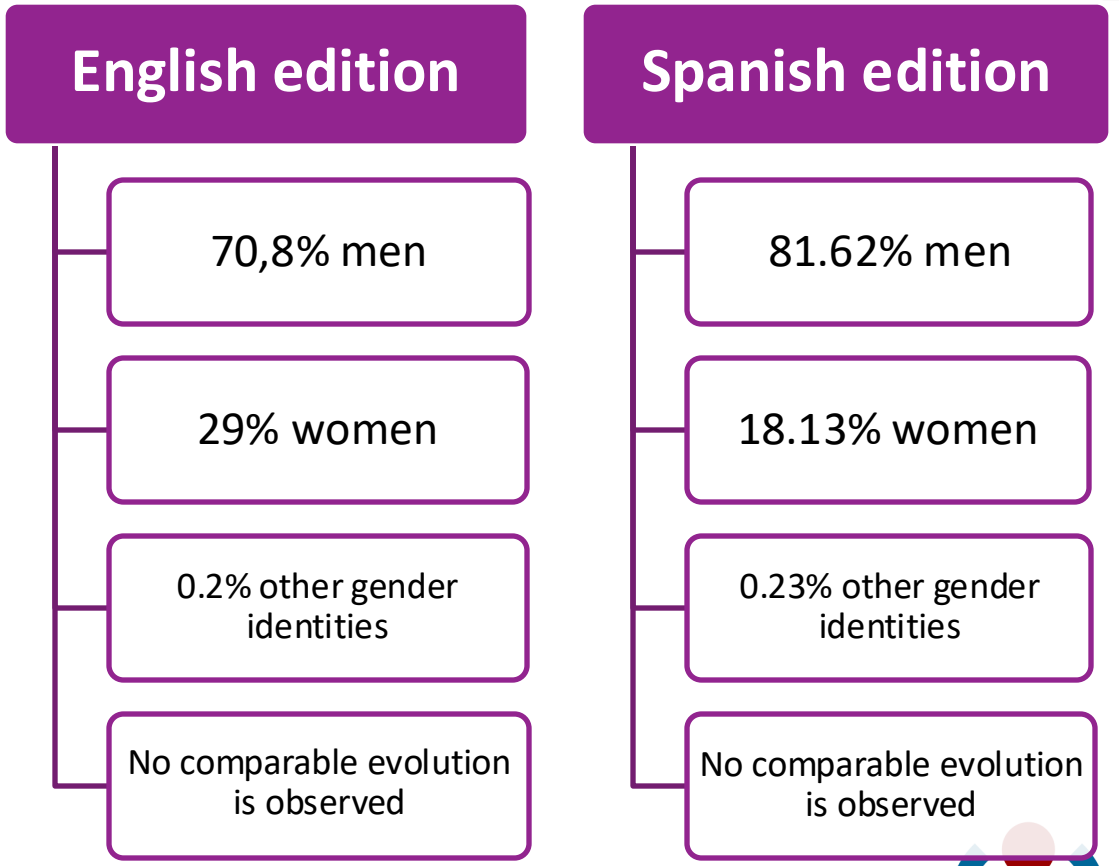
Wikimedia Commons: July 4th USA Flag. Photo by US Embassy Accra (Ghana).

c)Specific results 4: front page content

- **Ethnicity P172**
    - **English edition:** The term ‘African American’ appears 288 times, sharing this frequency with 183 other values. The top 20 values in this dataset together account for 25.9% of all appearances. In contrast, the representation of Europe is comparatively low, at 8.3%, with 33 values, while that of Latin America is even lower, at 0.96%, with 17 appearances out of a total of 1,761.
- These results suggest that, in the context of Wikipedia, ‘white-European-Caucasian’ may be perceived as the **default or absence of an ethnic group**.
- This discrepancy in representation raises questions about the extent to which Wikipedia's coverage of ethnic groups is equitable and reflective of the world's diversity.
- Mainstream data is not gathered, while non-normative identities are gathered.

c)Specific results4 : front page content

- **Sex or gender (P21):**
  - English Wikipedia:
    - 2017 was the year with the most balanced gender representation, with a gap of 630 values.
    - By 2022 the gap had widened to 1,191, almost doubling the difference
    - Despite global initiatives to increase the visibility of women on Wikipedia, this trend has not been reversed, highlighting the persistence of gender disparities.
  - Spanish Wikipedia:
    - Data indicates that the gender disparity is even more pronounced in the Spanish edition.



c)Specific results: front page 5

Sex or gender (P21):

- English Wikipedia:
  - 2017 was the year with the most balanced gender representation, with a gap of 630 values.
  - By 2022 the gap had widened to 1,191, almost doubling the difference
  - Despite global initiatives to increase the visibility of women on Wikipedia, this trend has not been reversed, highlighting the persistence of gender disparities.

- Spanish Wikipedia:
  - Data indicates that the gender disparity is even more pronounced in the Spanish edition, with a higher percentage of men and a lower representation of women and other gender identities.

Catalan Wikipedia achieved gender parity in 2023 in its cover page!

Social implications: Cover Women results

- **Impact on Knowledge Equity**:
  - Wikipedia's content perpetuates societal biases.
  - Marginalized groups lack visibility and representation.

- **Importance of Wikidata**:
  - Inclusion of diverse gender identities in Wikidata can influence Wikipedia's taxonomy.

Conclusions

- The Cover Women project highlights the **structural biases in Wikipedia's main page content:**
  - Wikipedia front pages in English and Spanish exhibit significant implicit biases, particularly regarding gender, ethnicity, language, and geographic representation.
  - Content prioritization reflects the personal interests of volunteer editors, leading to an overrepresentation of politicians and writers while underrepresenting women, minority ethnic groups, unconventional professions, and less prominent regions like Africa and Latin America.
- **Barriers to equity**: The editorial guidelines aim for neutrality and quality but fail to address systemic inequities effectively. While some guidelines encourage prioritizing diverse content, their implementation is inconsistent and limited in impact
- **Community power dynamics**: Decisions on content are heavily influenced by experienced editors and administrators. This hierarchical structure often reinforces the status quo, discouraging new, diverse contributors and maintaining gender and cultural gaps.
- **Challenges of representation**: Reliance on traditional notions of "notability" disadvantages underrepresented groups due to a lack of media coverage about them, further perpetuating biases in Wikipedia content.

Recommendations and librarians' role 1

- **Librarians and information professionals play a pivotal role in fostering inclusive digital knowledge spaces.**
  - Simplify editorial guidelines to enhance accessibility for new editors.
  - Revise the "notability" criterion to account for systemic disparities in media representation.
  - Promote initiatives that actively address the gender gap and other biases.
  - Consider the potential of AI and bots to democratize editorial processes while monitoring their biases.

- **Proposed Actions**:
  - Collaborate with Wikimedia projects to improve editorial policies.
  - Create templates to encourage balanced representation on Wikipedia's main page.
  - Facilitate discussions on inclusion in the Wikimedia community.

- **How Librarians can help**:
  - Assist in developing procedures for managing and achieving homepage content.
  - Promote training and workshops to engage diverse contributors.
  - Use data literacy skills to identify and address biases in featured content.
  - Advocate for diverse representation in Wikidata entries.

Recommendations and librarians' role 2

- **Collaborate in WIKIDATA:** Wikidata is a critical tool for reducing these gaps by expanding recognition of diverse identities.
  - Advocate for diverse representation in Wikidata entries.
  - Promote training and workshops to engage diverse contributors.
  - Use data literacy skills to identify and address biases in featured content.

- **Help overcome limitations of WIKIDATA**:
  - **Confusion Between Sex and Gender**: A recurring issue in Wikidata is the interchangeable use of property P21 ("sex or gender") to refer to both biological sex and gender identity. This creates ambiguities and makes it difficult to conduct precise analyses on gender.

  - **Amplification of biases on Wikipedia**: Since Wikidata is used as a data source for many functions on Wikipedia, any bias in Wikidata's information can amplify the issues of representation and diversity observed on Wikipedia's front pages. This creates a vicious cycle where already underrepresented groups remain invisible.

Co-creation of editorial guidelines

- **Collaborate in co-creation sessions:**
  - Organize them. Cover women project can provide the explanation on how to implement the session and the support materials.
  - OR
  - Just bring volunteers and Cover Women project can do the rest.

- **Expected outcomes:**
  - Prioritazion of criterias and editorial policies for each section of the Front page.
  - Draft document of an editorial policy to be discussed for implementing in the Front page.

- **Pilots in the Wikipedia in Catalan, Spanish forthcoming... anybody else?**

Thanks for your attention!

More information: www.ub.edu/wikiwomen

nferranf@ub.edu

laurafernandez@ub.edu

miquel.centelles@ub.edu

Research Grant
G-RS-2402-15223

UNIVERSITAT DE BARCELONA