Journal of Internet Services and Applications

RESEARCH Open Access

The Wikipedia Diversity Observatory: helping communities to bridge content gaps through interactive interfaces

Marc Miquel-Ribé* and David Laniado

* Correspondence: marc.miquel@ce.eurecat.org
Eurecat, Centre Tecnològic de Catalunya, Barcelona, Spain

Abstract

In this paper, we present the Wikipedia Diversity Observatory, a project aimed to increase diversity within Wikipedia content. The project provides dashboards with visualizations and tools which show content gaps in terms of imbalances in the coverage of topics, and of concepts that are not shared across Wikipedia language editions. The dashboards are built on datasets generated for each of the more than 300 existing language editions, with features that label each article according to geography, gender and other categories relevant to overall content diversity. Through various examples, we show how the tools encourage and help editors to bridge the gaps in Wikipedia content. Finally, we discuss the project's impact on the communities and implications for the Wikimedia movement in a moment in which covering diversity is considered strategic.

Keywords: Diversity, Culture gap, Gender gap, Online collaboration, Wikipedia, Digital humanities, Data visualization

1 Introduction

Wikipedia is among the most extensive information repositories on the Internet that are multilingual and created through a collaborative effort. Its prime objective is to "give free access to the sum of all human knowledge," Consequently, it exists in as many as 309 languages distributed worldwide.

Even though the language communities make the projects grow on a constant basis, the content does not represent the existing diversity in terms of peoples, places, and cultures; furthermore, there is a gap between Wikipedia language editions and articles often are not shared between them, sometimes remaining even exclusive to one language edition [1].

Ever since 2006, Wikipedians created a page to denounce the "imbalanced coverage of subjects and perspectives on the encyclopaedia" and called it systemic bias. As stated in it, imbalances in the representation of topics and content gaps were primarily associated with "most editors' shared social and cultural characteristics." This hypothesis has been confirmed over the years as several studies proved the relationship between

content imbalances and the lack of diversity in the community in terms of geography or gender [2, 3] along with the implicit biases [4].

More largely, the topical coverage of Wikipedia is the result of the interplay of factors including editors' motivation [5], quality content preferences [6], personal ideologies [7] and the policies that limit the acceptability of content [8, 9], among others. While the contextualisation of content in a collaboration space like Wikipedia is inevitable, the extent of imbalances and content gaps implies that Wikipedia is reflecting the structural and representative inequalities of our world, and given its influence, it can amplify and deepen them [2, 10, 11].

From an academic perspective, Wikipedia offers the opportunity to understand how contextual factors and biases shape the content. Building on this approach, we present the Wikipedia Diversity Observatory, a project aimed to increase diversity within Wikipedia content. Extending our previous work on identifying content related to a given cultural context [12–14], we developed a broader framework to characterize content diversity along different dimensions, including geography, gender, LGBT+, ethnic groups, and religious groups. On top of this computational approach, we developed a set of interactive dashboards to assist the communities in assessing the gaps and imbalances in the content, and identifying actions to reduce them.

The Wikipedia Diversity Observatory is an in-development system that is built through different iterations, each of them motivated by the period feedback received by the authors at different Wikimedia community events. The system is open for observation and invites everyone to engage in the development, in line with the ethos of the Wikimedia Movement.

The purpose of this paper is to disseminate the project and its approach, results, and impact on the communities. The main contributions are the following:

- proposing an iterative, open research approach to work with and for the Wikimedia communities, to create solutions aimed at improving content diversity in Wikipedia;
- extending our computational method to identify content gaps with respect to underrepresented categories along different dimensions, including gender, geography, ethnicity, and in all Wikipedia language editions;
- presenting a set of dashboards and tools to make the data and results easily explorable and to provide actionable knowledge for the communities;
- showing through selected use cases how the visualizations and tools proposed can be used in the daily work of the editors;
- documenting how input and feedback from the communities were integrated at different project development steps and the impact of the proposed dashboards on Wikimedia events and contests.

In the following section we provide background for our work, explaining what we know about content gaps both from an academic and a Wikimedia community perspective (Section 2), then we describe the approach followed to create the dashboards (Section 3), and illustrate through use cases the visualizations (Section 4) and tools (Section 5) created. We further discuss community engagement in the iterative

processes through which we developed our research and tools (Section 6). Finally, we draw conclusions and lines for future research (Section 7).

2 Background

2.1 Knowledge representation and content gaps

Several studies showed that geographical factors influence the topical distribution of content in Wikipedia language editions [13–20]. The **geography gap** means that some areas of the world are poorly represented in Wikipedia, Wikidata and the other sister projects [21].

Editors tend to edit about places near where they are editing [22, 23]. However, the lack of connectivity (i.e. digital divide) and other factors prevent billions of people to contribute to Wikipedia, or even to access it, which creates an uneven representation of the world [24, 25].

In fact, the most active language editions tend to represent extensively the cultural context where the language is spoken, dedicating articles to a variety of topics (i.e. their places, traditions, language, art, popular culture, agriculture, biographies, etc.), but fail to ensure a minimum coverage of the other languages' related cultural and geographical context.

The content devoted by each language edition to its corresponding territories and culture has been defined as Cultural Context Content (CCC) [13]. Such content occupies about a quarter of the first 40 language editions in the number of articles, with cases in which it occupies over 44.2% (English) and others with as little as 9.0% (Dutch) [13].

Far from being a one-time event, the creation of CCC is sustained over time. Editors create it regularly and often refer to it as "local content," in opposition to the articles that are expected to be in every Wikipedia language edition as notable global knowledge. Most language editions only cover local content related to areas that are geographically close, and still do it very partially (the **culture gap**) [13].

However, cultural aspects are also reflected in content at the in-article level, overrepresenting some perspectives and overlooking others in biographies, historical events, or politics especially depending on the Wikipedia language edition [26–28]. This also happens in the coverage of images: Ahmed and Poulter [29] compared the coverage of visual arts in different projects, including Wikimedia Commons, and found that European language editions of Wikipedia are generally more "Western" in their coverage and Asian languages more "global."

These gaps are also relevant because they may stem from disputes between national identities and from silencing specific points of view in a Wikipedia language. It is, for example, the case of the edit wars which took place in the renaming of the article of the river Ganga/Ganges in the English Wikipedia; clearly, the conflict reproduced postcolonial politics, and the outcome was only a matter of persistence and numerical strength [30].

The representation of ethnic minorities in Wikipedia is also insufficient. Some studies show the importance of engaging **indigenous groups** in Wikipedia, and at the same time, its implicit challenges both regarding the context and even the definition of knowledge [31, 32]. Initiatives and global campaigns like "Decolonize the Internet" are looking into practices that can support marginalized communities to begin centring their knowledge online [2].

Possibly the most known kind of content gap is the **gender gap**.[1] Even though it is often depicted as a lower percentage of women in biographies [3, 34–36], other authors have detected that it also affects the extent of meta-data, the language used to represent women, and the visibility in the Wikipedia network structure [37, 38]. Less discussed is the **LGBT+ gap**, which has been described in relation to topics as diverse as the fight for rights, the culture, and the expression of sexuality in biographies. Wexelbaum et al. [39] have studied the strategies Wikipedians deploy to bridge the LGBT+ gap and the importance of Wikipedia as the most useful resource on the Internet to increase the visibility of the corresponding content in many contexts.

2.2 Strategic direction and content diversity

In the past years, there has been a growing awareness of the need for increased content diversity in the communities. The Wikimedia Foundation initiated a Movement Strategy Process to understand the future priorities of the Wikimedia Movement. One of the resulting two goals set for the 2030 horizon is to reach "knowledge equity",[2] which implies to "counteract structural inequalities to ensure a just representation of knowledge and people in the Wikimedia movement."

Community initiatives that go from conferences and global campaigns[3] to online contests have proliferated to coordinate efforts to bridge different kinds of gaps. In 2018, Wikimania, the annual international conference, was held in South Africa with the theme "Bridging the Knowledge Gaps – The Ubuntu Way Forward"[4] to put emphasis on Africa's under-representation.

In parallel to these initiatives, tools to monitor different kinds of gaps in Wikipedia content have started to be developed, especially for the gender gap [35]. Other content gaps take longer to be measured because of their complexity, and gap monitoring tools are not available yet. At the same time, some efforts are also being deployed in order to create a general conceptual framework to classify gaps in Wikipedia both in the readers, editors, and articles [40].

As of now, no project has aimed at both showing the gaps and providing suggestions to bridge them so that editors can immediately act. Hence, we present the Wikipedia Diversity Observatory as a project that creates a space for both scholars and Wikipedia editors to identify and bridge content gaps. The **main objective** of this project is to address the need to measure, characterize and monitor the coverage of underrepresented groups of people, places, and cultures, and finally provide suggestions of top priority articles to be created in specific languages in order to bridge content gaps.

We share the experience of having a unified site for all Wikipedia language editions[5] based on a framework created to collect, process, expose and visualize data, providing the code released under open source license.[6] The project aims to integrate easy-to-use

[1] It should be noted that the term "gender gap" is sometimes used to refer to gender imbalance in the editor community, where the proportion of women is estimated to be around 15% [33]. As in this study we focus on imbalances in content coverage, in the following we will refer to gender gap in the content, and specifically to imbalances in the coverage of biographies.
[2] https://meta.wikimedia.org/wiki/Strategy/Wikimedia_movement/2018-20
[3] https://whoseknowledge.org
[4] https://blog.wikimedia.org/2018/02/05/wikimania-cape-town-ubuntu/
[5] The project visualizations and tools are available at http://wdo.wmcloud.org/.
[6] The project code is available at https://github.com/marcmiquel/wcdo.

dashboards into the community's daily activities to raise awareness by showing gaps and proposing specific solutions.

3 Approach

3.1 Open research process

The Wikipedia Diversity Observatory[7] originated as a community initiative supported by a project grant to provide both valuable metrics to understand the current gaps in any Wikipedia edition and actionable results that can help editors bridge them as part of their wiki activities. This differentiates its approach from many projects created in a company or organization with a "target user" and a user research approach. In our case, the project proposal was approved thanks to support and endorsements from the communities,[8] with the promise of creating a space for collaboration and research on the content gaps.

As stated in its Wikimedia Meta page, the Diversity Observatory page[9] "is a joint space for editors, researchers and all sort of contributors to study and fight against the content gaps." On the one hand, this means that, while the authors of this paper have been behind the research and development, many volunteers from different language communities are engaging in giving feedback on general aspects of content diversity, and more specifically, on the use of and experience with the tools created. On the other hand, this also means that all the content generated, code created, and data stored are made available along with detailed instructions for anyone to be able to get involved[10] and engage in tasks they would like to carry out.

We chose to develop this project following an open research model [41, 42] because it is the most convenient approach to engage with the Wikimedia communities and amplify its goals. These are: "raising awareness on Wikipedia's current state of diversity according to specific topics and categories" and "providing datasets, visualizations, and tools to improve on it." Each of them is tackled by an iterative process (Fig. 1) with various phases and subphases that are encompassed in each iteration.

The first phase, **dataset generation and analysis**, is primary as it focuses on obtaining and understanding the data related to diversity-related topics. Its research findings are used to create the tools and visualizations in the second phase, **dashboards development**. Then, **community engagement** is a central phase dedicated to confirming the value of the research findings, generating new topics and questions, and understanding the needs of users when using the tools and listening to their specific requests.

While this is presented as an iterative process, prior to the first iteration we need to identify the different relevant communities and subcommunities and the types of gaps they are interested in. We explain this in subsection Community Exploration (3.2), then in subsection Dataset Generation and Analysis (3.3) we detail the main steps to generate data, and finally, in subsection Dashboard Development (3.4) we describe the visualizations created to understand the extent of the gaps, and the tools that provide lists of articles to bridge them.

[7] https://meta.wikimedia.org/wiki/Wikipedia_Diversity_Observatory
[8] https://meta.wikimedia.org/wiki/Grants:Project/Wikipedia_Cultural_Diversity_Observatory_(WCDO)
[9] https://meta.wikimedia.org/wiki/Wikipedia_Diversity_Observatory
[10] https://meta.wikimedia.org/wiki/Wikipedia_Diversity_Observatory#Get_involved

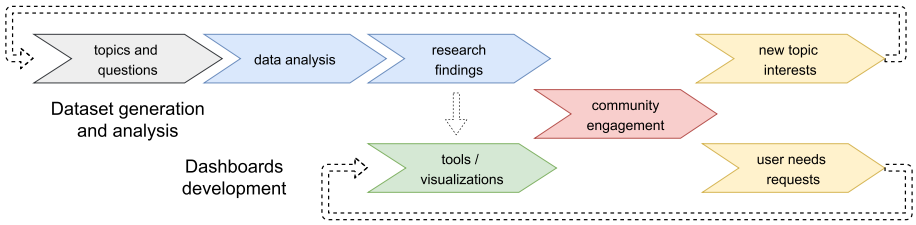

Fig. 1 Wikipedia Diversity Observatory open research process phases

3.2 Community exploration

As a first step, we needed to identify the Wikipedia communities active on each category or topic related to possible content gaps in one or more language editions. This would serve two purposes.

First, we would understand more about the topics that might be worth analysing in the future, among those which are not covered spontaneously. While the Diversity Observatory started with a specific focus on the culture gap, learning about community concerns for other content gaps led us to extend the scope of the project.

Second, we would recognize the specific communities or groups already working on content gaps, understand their concerns and be able to let them know about the project and contact them at later stages. As we explain in Section 6, community engagement is an essential step in the development of the project, considering that the Diversity Observatory is a space for both researchers and activists and the need for input from diverse Wikipedians is valuable to define the problems, refine the analyses, and create the visualizations and tools that suit them.

For this initial community exploration, we took a look at the Wikimedia project "Meta-Wiki".[11] This is the global community site for the movement projects for coordination and documentation. There we found the pages for the different projects and user groups aimed at bridging specific gaps. We looked specifically at the editors' level of engagement, the way they coordinate, and the terminology and categories they employ to refer to the gap.

Considering only the formal groups (user groups) listed in the Wikimedia movement affiliates page,[12] we found 6 groups with gender as the main theme, 1 dedicated to LGBT+, and 1 to Indigenous Languages. Many others address geographical and culture gap as part of their scope, focusing on specific geographical and cultural contexts. Some others focus on general encyclopedic topics such as medicine, maths, and cartography.

Gender groups like Les sans pageES,[13] WikiWomen,[14] and Art+Feminism[15] are aimed at bridging the gender gap, and centre their efforts on creating biographies of women, among other activities (Fig. 2). With regard to the geography gap, we realized that it is never mentioned directly nor addressed as a whole by user groups, but segmented into specific territories. For example, increasing the quantity and quality of the articles about Africa are the goals of the group and project named WikiAfrica.[16]

[11] https://meta.wikimedia.org/
[12] https://meta.wikimedia.org/wiki/Wikimedia_movement_affiliates
[13] https://meta.wikimedia.org/wiki/Les_sans_pagEs
[14] https://meta.wikimedia.org/wiki/WikiWomen%27s_User_Group
[15] https://meta.wikimedia.org/wiki/Art%2BFeminism_User_Group
[16] https://meta.wikimedia.org/wiki/WikiAfrica

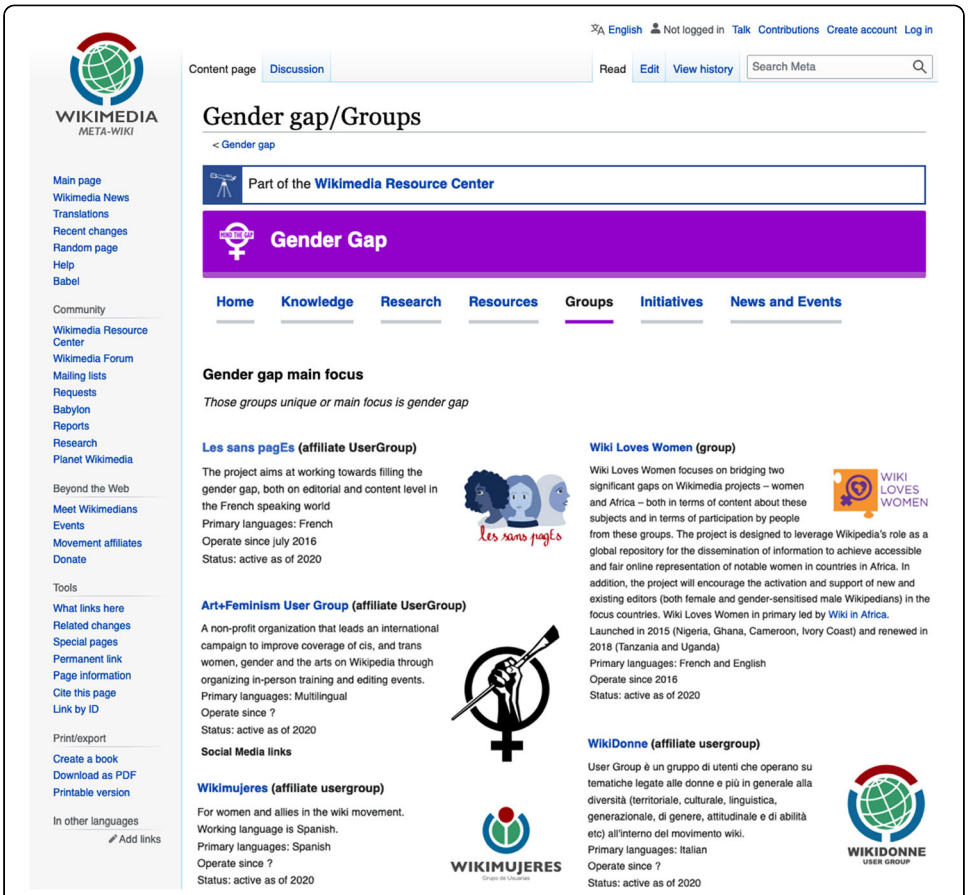

Fig. 2 Wikimedia user groups whose unique or primary focus is the gender gap

In fact, we also realized that when groups focus on a particular territory, they also tend to include articles about their cultural context content and address the culture gap and the geography gap at the same time. For example, the global contest "Asian Month[17]" is celebrated in 60 Wikipedia language editions, and it sets the rule that all the articles created within the event should concern an Asian country. In the resulting lists of articles created after this year's edition of the contest, we see biographies, temples, towns, music bands, traditions, among many other topics. The same thing happens in the CEE Spring contest, in which the language editions corresponding to the Central and Eastern European countries create articles about each other's cultural context. Sometimes, the gender-focused groups of editors participate in these contests and arrange parallel more specific competitions, such as to create 100 biographies of women from Asia.

Similarly, the gaps in articles about biographies and culture of ethnic and indigenous groups are usually addressed by the groups geolocated in an area that includes the corresponding territories, sometimes being this their entire scope (e.g., Wikimedians of North American Indigenous Languages User Group[18]), other times as a part of their activity (e.g. Wikimedia Canada[19]).

[17] https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_Asian_Month
[18] https://meta.wikimedia.org/wiki/Wikimedians_of_North_American_Indigenous_Languages_User_Group
[19] https://ca.wikimedia.org/wiki/Indigenous_communities_outreach

Concerning LGBT+ related content, there are both a user group[20] and a portal[21] in 39 language editions aimed at classifying all the articles and projects related to LGBT+, including biographies, identities, culture, among others.

3.3 Datasets generation and analysis

Once we had a good understanding of the main gaps communities have detected and are organized to work on, we needed to identify and store the relation between each of the categories relevant for diversity and the articles.

As a first step, we designed a method in which each article is characterized according to features that can determine whether it belongs to a relevant category for diversity (culture, gender, place, etc.). Categories like gender, sexual orientation, religion or ethnic origin are straightforward, as they can be traced to Wikidata semantic relations structured as properties and items. For example, Elton John in Wikidata has the property sex or gender assigned to male and sexual orientation to homosexuality.

Instead, associating an article to a language's cultural context (i.e. as part of the local content) requires a more sophisticated method. In this case, we do not only use Wikidata properties and Qitems, but also information stored in the articles, considering that in many languages the information available is richer than in Wikidata - especially for articles related to the cultural context. We use a variety of features based on the article title, category, and links graph structure, among others, to label each article according to the possible relationship with territories where the language is spoken and to the peoples that inhabit them. A detailed description of these features is presented in [13].

As a second step for the cultural context content, we introduce all of these features into a machine learning classifier to obtain the final selection of articles belonging to a language's context, following the approach described in [14]:

- **Classifier:** Random forest classifier with negative sampling (as we did not have a representative set of negative items, the classifier was trained to distinguish positive from random articles).
- **Training data:** Groundtruth of articles having features that strongly and reliably associate them to the language's cultural context (e.g. geolocation, keywords in the article title, strong Wikidata properties like "country of birth").
- **Testing data:** All articles having at least some weak features associating them to the language's cultural content.

We performed a manual assessment of the results for 10 diverse language editions, retrieving 200 random articles from each of them (100 classified as positive and 100 as negative by the algorithm). For all 10 languages, precision was between 93% and 100%, and recall between 94% and 100%.

We employ a similar approach for topics related to LGBT+ and topics related to ethnic groups. The resulting datasets are available in different formats (e.g., Sqlite3 and CSV)[22,23] and are computed regularly.

[20] https://meta.wikimedia.org/wiki/Wikimedia_LGBT+
[21] https://en.wikipedia.org/wiki/Portal:LGBT
[22] https://wdo.wmcloud.org/databases
[23] https://doi.org/10.6084/m9.Fig.share.7039514.v3

As a third step, based on the dataset produced for each Wikipedia language edition, we created a database[24] with **monthly data and metrics on content diversity and gaps**.[25] This consists of some basic statistics for groups of articles representing content associated with a language, a territory at different levels of granularity (e.g., Europe, Southern-Europe, Italy), or other categories relevant for diversity in the overall content (e.g., gender), and their intersections with one another and with larger groups of articles (e.g., an entire language edition, or articles created during the past month).

3.4 Dashboards development

Finally, building on the database with all the articles related to each category relevant to diversity, and on the database with the statistics, created dashboards with visualizations and tools. These are updated on a regular basis to allow for comparison of the extent and coverage of specific groups of articles (e.g., content related to the culture associated with a given language or territory, articles geolocated within a given region, or biographies of people having specific characteristics such as gender, ethnic group, religion or sexual orientation) across language editions. While the visualizations allow one to monitor the progress in bridging the gaps between language editions, the tools provide specific lists of articles and other content suggestions to foster the creation, improvement, and exchange of content.

Visualizations are pages dedicated to showing different kinds of content gaps and their measure across language editions:

- **Culture Gap**[26,27] illustrates how well each Wikipedia language edition covers the CCC from the other language editions and how well each Wikipedia language edition's CCC articles are spread across languages.
- **Geographic Gap**[28] illustrates how well each Wikipedia language edition covers all the existing geolocated articles categorized according to the different geographical entities (country, subregion, and world region).
- **Gender Gap**[29] illustrates the gender gap in Wikipedia language editions content at an article level, taking the number of biographies as a proxy.
- **Ethnic Groups Gap**[30] illustrates coverage of ethnic groups Wikipedia language editions content based on biographies of people belonging to different groups and topics that relate to the cultural context associated with each ethnic group.
- **Religious Groups Gap**[31] illustrates the gap in coverage of religious groups across Wikipedia language editions, taking into account people's biographies with a religious group affiliation.
- **Last Month Pageviews**[32] illustrates the distribution of pageviews in the previous month for the different categories of diversity in each Wikipedia language edition.

[24] https://wdo.wmcloud.org/databases/stats_production.db
[25] https://meta.wikimedia.org/wiki/Wikipedia_Diversity_Observatory/Sets_intersections_and_increments
[26] https://wdo.wmcloud.org/ccc_coverage/
[27] https://wdo.wmcloud.org/ccc_spread/
[28] https://wdo.wmcloud.org/geography_gap/
[29] https://wdo.wmcloud.org/gender_gap/
[30] https://wdo.wmcloud.org/ethnic_groups_gap/
[31] https://wdo.wmcloud.org/religious_groups_gap/
[32] https://wdo.wmcloud.org/last_month_pageviews/

- **Diversity Over Time**[33] illustrates the creation of content belonging to the different categories of diversity over time. It depicts both the accumulated articles and the new articles created on a monthly basis.
- **Recent Changes Diversity**[34] shows the list of recent changes,[35] the most recent edits made to pages, according to the different categories relevant to diversity.
- **Tools** are pages that contain actionable knowledge to help editors bridge the gaps:
- **Top CCC Diversity Lists**[36] allows retrieving lists of top priority articles about any language edition's related cultural context or any other diversity category and checking its availability in any Wikipedia language editions to identify missing articles to be created.
- **LGBT+ Articles**[37] allows retrieving LGBT+ related articles from any Wikipedia language edition and check their availability in a specific Wikipedia.
- **Ethnic Groups Articles**[38] allows retrieving articles related to ethnic groups from any Wikipedia language edition and check their availability in a specific Wikipedia.
- **Time Articles**[39] allows retrieving articles with time properties in Wikidata about all kinds of topics and check their availability in a specific Wikipedia. You can search any topic and filter by particular features, e.g., articles with most interwiki from specific centuries.
- **Common CCC**[40] allows consulting a list of articles related to more than one language CCC at the same time. In other words, it allows one to retrieve articles that could belong to more than one cultural context.
- **Missing CCC**[41] allows consulting a list of articles that could and might need to exist or be extended in a language CCC, as they are part of that language's context, and instead, they only exist in other Wikipedia language editions.
- **Incomplete CCC**[42] allows assessing the completeness of a list of articles from a language edition introduced manually by the user, or of a "Top CCC Diversity List" by comparing it to the versions of the articles in other language editions.
- **Search CCC**[43] allows searching for articles in a Wikipedia language edition according to different categories relevant to diversity, also through the use of keywords, and see their availability in other language editions.
- **Visual CCC**[44] allows searching for missing images (visual gaps) in the articles of a Top CCC list or a list of articles specified by the user, which are used in the other language editions' versions of the article.

4 Visualizations

Since the Diversity Observatory database categorizes all the Wikipedia language editions articles according to different types of gaps, we can visualize them both transversely and longitudinally.

[33] https://wdo.wmcloud.org/diversity_over_time/
[34] https://wdo.wmcloud.org/recent_changes_diversity/
[35] https://en.wikipedia.org/wiki/Help:Recent_changes
[36] https://wdo.wmcloud.org/top_ccc_articles
[37] https://wdo.wmcloud.org/lgbt+_articles
[38] https://wdo.wmcloud.org/ethnic_groups_articles/
[39] https://wdo.wmcloud.org/time_articles/
[40] https://wdo.wmcloud.org/common_ccc_articles
[41] https://wdo.wmcloud.org/missing_ccc_articles
[42] https://wdo.wmcloud.org/incomplete_ccc_articles
[43] https://wdo.wmcloud.org/search_ccc_articles
[44] https://wdo.wmcloud.org/visual_ccc_articles

4.1 Culture gap coverage and spread

For example, in the dashboards dedicated to the Culture Gap (CCC Coverage and CCC Spread), we can see, on the one hand, how well each Wikipedia language edition covers the CCC of the other language editions, and on the other hand, the extent of all language editions' CCC in every Wikipedia language edition. For example, in Fig. 3, we can see the number of articles from Javanese CCC that exist in other language editions. It appears that Indonesian (id) and Dutch (nl) language editions are the ones that best cover Javanese CCC (the first one with more than 32 k articles, and 83.4% of the entire Javanese CCC, and the latter one with over 17 k articles, which equates to a 45.7%). The results from these two languages can be explained by the linguistic overlap (Java is a region of Indonesia, where both Javanese and Indonesian are spoken) and the colonial power that Netherlands exerted in Indonesia until 1949. The following languages are Malay, which could be explained by the millions of Javanese descendants who were born or immigrated to Malaysia, and English, the international language that tends to have the highest coverage for most topics, including other languages' CCC. The colour of the bars account for the different overall sizes of the corresponding languages: English is depicted in yellow as it contains more than 6 million articles; Dutch in purple as it includes around 2 million articles; Indonesian and Malay are depicted in blue as they have a much lower overall number of articles (555 k and 345 k articles, respectively). In this way, while the graphic is not normalized and shows the absolute number of articles from Javanese CCC covered by each language edition, colours help the reader put this quantity in relation to the overall size of a language edition.

In Fig. 4 we can see a treemap graph which shows the extent occupied by CCC from each language in the Dutch (left) and Indonesian (right) Wikipedia (considering only articles belonging to CCC from some language edition). In the Dutch Wikipedia, its own CCC takes 6.62%, English CCC 7.72% and Javanese CCC only 0.88%. In the Indonesian Wikipedia, we see that the Indonesian CCC takes 21.68%, the English 11.4%, and the Javanese a 5.94%. Therefore, we can see that the extent CCC from other language editions takes in every language depends on the proximity, but also on the overall size of the language and the original number of articles in that language CCC. While the treemap graph allows one to intuitively see how much the representation of each cultural context takes in a given Wikipedia (i.e. it is based on the "extent" percentage), the coverage percentage is also reported in each cell, to indicate the proportion of a given cultural context in content. For example, in Fig. 4 we can see that the English CCC is the largest in the Dutch Wikipedia, yet only 7.5% of it is covered.

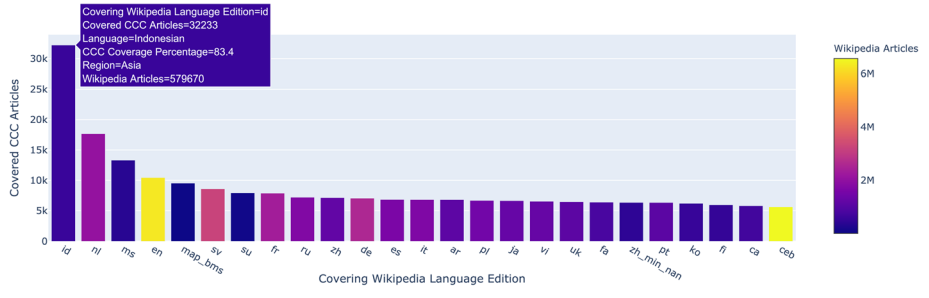

Fig. 3 Interactive barchart showing the number of articles from Javanese CCC that exist in other Wikipedia language editions. The colour indicates the total number of articles in those language editions

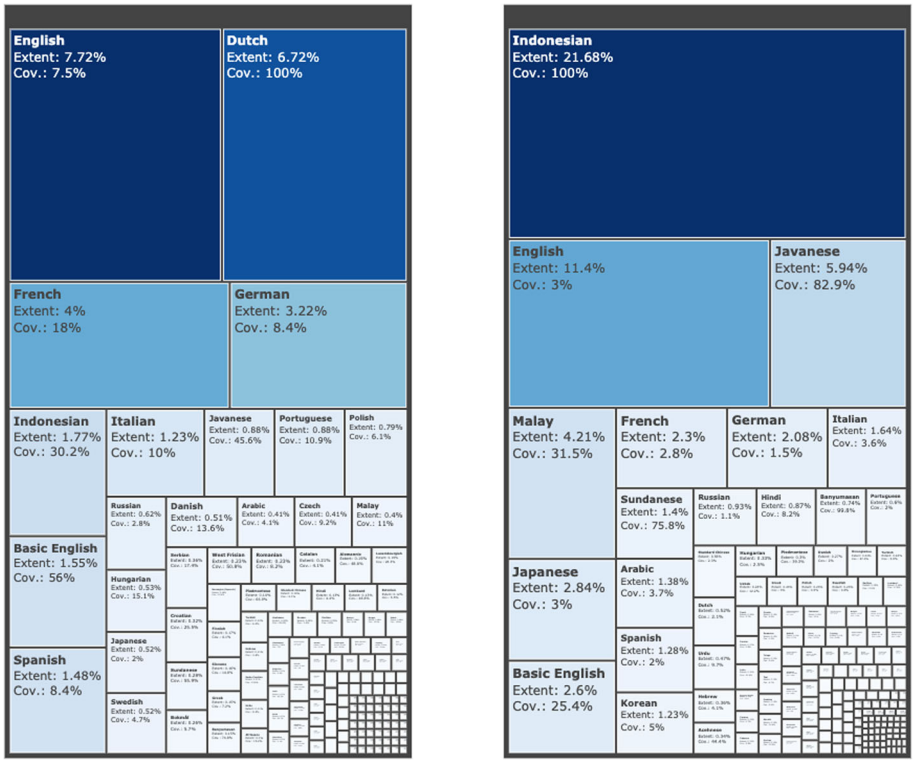

Fig. 4 Interactive treemap shows CCC's extent and coverage from other languages in Dutch (left) and Indonesian (right) Wikipedia language editions

4.2 Geography gap over time

In the dashboard Diversity Over Time, we can see the creation of articles for one or more diversity categories in multiple language editions over time. We can choose whether to compare a specific entity (geographical entity like continent or subcontinent, gender or language culture) and a group of language editions or a group of entities for a single Wikipedia language edition. In Fig. 5, we see the creation of articles geolocated in Sub-Saharan Africa over the past 12 years in five of the largest Wikipedia language editions. While on the left graph, we see the growth in the absolute number of geolocated articles; on the right graph, we see the relative value, normalized by the total number of geolocated articles in each language edition. We can see that Sub-Saharan Africa occupies a maximum of 1.2% of the articles with a geolocation tag in these languages editions, with the highest value for the French Wikipedia. It is important to note that despite having dedicated the Wikimania 2018 conference to the lack of articles related to Africa, we hardly see an impact on geolocated article creation, as the percentages remain stable.

4.3 Gender gap in pageviews

In the dashboard Last Month Pageviews, we can see a series of comparisons between the distribution of articles and pageviews for the diversity-related categories geography, CCC and gender. This way, we can understand whether a specific category receives a higher proportion of pageviews than the proportion of articles it occupies in the

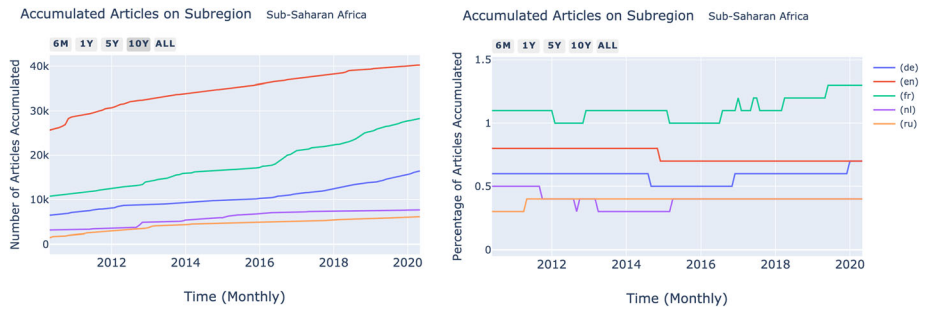

Fig. 5 Monthly evolution of the overall number of articles geolocated in Sub-Saharan Africa in the German, English, French, Dutch and Russian Wikipedia, in absolute numbers (left) and normalized by all geolocated articles (right)

language edition. This is precisely the case for the gender gap (Fig. 6). These stacked bars show the gender gap in biographies, first in the number of articles and second in the number of pageviews these have received. For this particular graph, we selected the first ten languages in the number of editors. We can see that in all cases the proportion of pageviews men receive is lower than the proportion of articles, which indicates that despite women have fewer biographies, they receive in average more pageviews (e.g. 81.4% for men biographies in English Wikipedia, and 69.4% for the pageviews they received).

5 Tools

While the visualizations help to depict the situation, the tools point out specific gaps and provide suggestions for editors to act on specific topics. We will illustrate two cases to show how dashboards can help bridge the culture gap.

5.1 Case 1: culture gap (Top CCC Diversity Lists): sharing the topics related to the language context across language editions

The Top CCC Diversity Lists[45] help editors discover valuable articles from each language's cultural context and immediately see their coverage by other language editions. Since the Top CCC Diversity Lists are associated to a language of origin, they address specifically the culture gap in the lack of articles about that language's related topics. Still, they can also be combined with other diversity categories like gender and geography, or even to topics like music, monuments, folk, among many others. There are lists for each of the topics of the various Wikimedia community programs and events that follow the pattern "Wiki Loves X", where X is Earth, Music, etc. Editors can retrieve articles specific to their interests, check their relevance, and choose an article to translate or adapt to another language edition.

In Fig. 7, we see the Top 500 articles from Yoruba CCC dedicated to women according to their number of edits (first column on the left) and their availability in Catalan Wikipedia depicted as red links or empty spaces for missing articles, blue links for articles which already exist (last column on the right). The remaining columns provide selectable article features such as the number of

[45] https://wdo.wmcloud.org/top_ccc_articles/

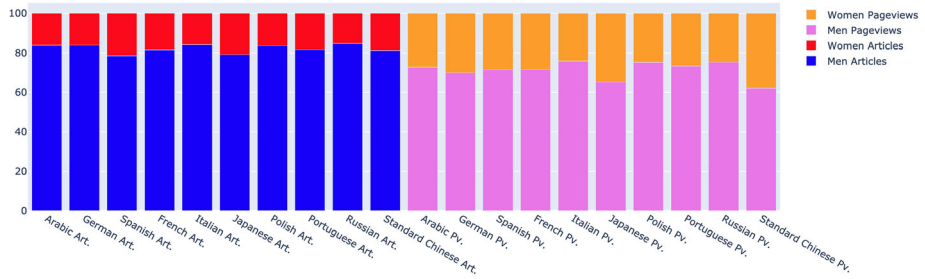

Fig. 6 Interactive stacked bars showing the gender gap in both the number of articles and in pageviews for the top 10 languages in terms of the number of editors

editors, length in bytes, number of languages in which it exists, among others. With more than 20 different Top CCC Diversity Lists for each language edition, any editor can verify the degree of coverage of the most relevant articles about every other language cultural context and some topics and bridge the gaps. Some additional dashboards show how well a specific language edition covers all the Top CCC Lists from every other language and how well their own lists are spread across other languages.

5.2 Case 2 culture gap (Missing CCC): representing the topics related to one's own language context using content from larger language editions

While the Top CCC Lists are useful to assess the coverage of diversity in the rest of the language editions, we observed that minor language editions do not sufficiently

Yoruba Top CCC articles list "Women" and its coverage by Catalan Wikipedia

Nº	Yoruba Article Title	Edits	Editors	Pageviews	Bytes	References	Creation Date	Related Languages	Catalan Article Title
1	Genevieve Nnaji	74	16	21	8.1k	12	2009-09-24	es, en, fr, it	Genevieve Nnaji (label)
2	Quincy Olasumbo Ayodele	36	3	0	8.3k	9	2016-07-12	en	Quincy Olasumbo Ayodele (translation)
3	Abiodun Olujimi	28	5	0	5.8k	3	2019-11-16	en	Abiodun Olujimi (translation)
4	Funmilayo Ransome-Kuti	25	6	1	5.8k	4	2009-12-10	es, en, fr, it	Funmilayo Ransome-Kuti
5	Ngozi Okonjo-Iweala	23	7	2	28.9k	89	2008-10-08	es, en, fr, it	Ngozi Okonjo-Iweala
6	Olúrẹ̀mí Ṣónáìyà	22	5	0	5.8k	5	2019-09-12	en, fr	Remi Sonaiya (translation)
7	Salawa Abeni	21	6	1	6.1k	9	2011-06-18	es, en	Salawa Abeni (translation)
8	Onyeka Onwenu	19	5	1	8.1k	28	2011-06-18	es, en	Onyeka Onwenu (translation)
9	Kemi Adeosun	19	5	0	6.6k	11	2017-06-08	en, fr	Kemi Adeosun (label)
10	Agbani Darego	17	5	0	1.1k	1	2009-12-19	es, en, fr, it	Agbani Darego (translation)
11	Chimamanda Ngozi Adichie	17	11	0	2.7k	2	2009-12-26	es, en, fr, it	Chimamanda Ngozi Adichie
12	Omotola Jalade Ekeinde	16	4	1	2.6k	1	2009-12-19	es, en, fr	Omotola Jalade Ekeinde (translation)
13	Babalola Chinedum Peace	15	3	0	4.5k	2	2018-10-30	en	Chinedum Peace Babalola
14	Hilda Dokubo	14	2	0	5.6k	7	2020-03-06	es, en, fr	
15	Joana Maduka	14	2	1	6.5k	19	2020-05-11	en	

Fig. 7 Interactive table showing a list of articles on women biographies related to Yoruba culture, sorted by the number of edits in the Yoruba Wikipedia. It shows the availability of each article in other language editions and points to the corresponding article in Catalan; when not existing, a red link points to a page to be created

represent their own cultural context, from their places to relevant public figures, their traditions, etc. This may typically derive from a small or scarcely active language edition community, and from contextual barriers to editing. In order to address this issue, the "Missing CCC" dashboard allows editors to search for articles that relate to their cultural context and exist only in larger language editions so that they can create the corresponding articles in their own language edition.

For example, the African language of Wolof is indigenous from Senegal, where it is the most spoken language, and it is also spoken in Mauritania. Surprisingly, Wolof Wikipedia has articles dedicated to the Scottish football coach Alex Ferguson, the American president Ronald Reagan, and the Italian theatre actress Anna Rita Del Piano, but none dedicated to the current president of Senegal and long-time politician Macky Sall. Although the existence of these articles may depend on the initiative of just some specialized editor, the contrast with having no article dedicated to such an important figure as the current president is striking.

When using the Missing CCC tool to search for articles from the Senegal context that are missing in Wolof Wikipedia, we find Macky Sall article in the 7th position of the results. This article exists in 48 language editions, including English (Fig. 8). Possibly, the creation of articles in the Wolof Wikipedia is partly following a Western view of which topics deserve to be included in an encyclopaedia, thus under-representing what may be relevant to Wolof readers. The results provided by the Missing CCC tool allow editors to identify articles that exist in other

Missing CCC Articles in Wolof Wikipedia on people

N°	Language	Title	Editors	Pageviews	Interwiki	Bytes	Lang Label	Qitem
1	en	Tacko Fall	118	53384	5	10.8k	fr Tacko Fall	Q18921708
2	en	Patrice Evra	1774	7346	63	133.0k	fr Patrice Évra	Q1916
3	en	Idrissa Gueye	212	5512	38	14.2k	fr Idrissa Gueye	Q46679
4	en	Patrick Vieira	1570	3271	61	83.1k	wo Patrick Vieira	Q46347
5	en	El Hadji Diouf	1592	2086	36	55.0k	fr El-Hadji Diouf	Q988626
6	en	Papiss Cissé	960	1400	34	34.1k	fr Papiss Cissé	Q309628
7	en	Macky Sall	152	1068	48	31.1k	fr Macky Sall	Q57438
8	en	Mame Biram Diouf	675	732	37	36.1k	fr Mame Biram Di	Q19051
9	en	Dame N'Doye	438	689	27	17.1k	fr Dame N'Doye	Q342501
10	en	Gorgui Dieng	126	632	21	19.3k	fr Gorgui Dieng	Q5586369
11	en	Diafra Sakho	177	604	29	29.1k	wo Diafra Sakho	Q1208103
12	en	Mor Thiam	90	582	3	3.6k	fr Mor Thiam	Q452985
13	en	Alfred N'Diaye	241	562	21	21.6k	fr Alfred N'Diaye	Q984689
14	en	Bambo Diaby	7	554	2	5.1k	fr Bambo Diaby	Q44079594
15	en	Lamine Sané	81	405	21	10.0k	wo Ludovic Sané	Q31803
16	en	DeSagana Diop	297	399	14	14.4k	fr DeSagana Diop	Q1179612
17	en	Hamady N'Diaye	88	399	6	13.1k	fr Hamady N'Diay	Q2632370
18	en	Papy Djilobodji	213	327	24	21.4k	fr Papy Djilobodji	Q3362956
19	en	Armand Traoré	593	299	36	33.4k	fr Armand Traoré	Q223138
20	en	Salif Diao	253	295	24	23.5k	fr Salif Diao	Q350785

Fig. 8 Interactive table showing a list of biographies related to Wolof culture, existing in other language editions and not in Wolof

language editions, sort them by relevance and identify those that may deserve more urgently being created.

5.3 Case 3 culture gap (Visual CCC): illustrating the topics related to one's own language context using pictures from other language editions

We have seen that while every language is more likely to represent its CCC better than the other language editions, this is not the case for many small Wikipedia language editions. The dashboard Visual CCC allows selecting a Top CCC list or providing a list of articles for a Wikipedia language edition and then obtaining a table with the most used images across the different language editions in which these articles exist. You can then filter out the images your language edition is already including for the list of articles and find those valuable images that are not used. For example, in Fig. 9, we can see a list of Top CCC articles for Albanian CCC sorted by the number of editors who edited them. The first result is Ismail Kadare, a renowned Albanian writer whose article exists in 59 language editions. In 23 language editions, they include his personal signature in the article, which is missing in the Albanian language version of the article. A picture of him reading a book is included in 54 language editions but missing in the Albanian one. This tool provides a direct and straightforward way to find visual gaps and bridge them.

Fig. 9 Interactive table showing a list of articles on Albanian Top CCC and the images that are missing in the Albanian version of the articles and exist in other language editions

6 Community engagement

In this section, we share the different "lessons learned" from the project development, specifically in regard to how we engaged with diverse Wikipedia communities to design solutions that addresses the needs of any Wikipedian willing to bridge the content gaps.

In the first subsection, we explain the general community needs we identified throughout the different interactions with the Wikimedia community members (6.1). In the second subsection, we detail the insights we obtained from each of the main iterations that have been fundamental to extend the data framework and build the tools and visualizations (6.2). Finally, in the third subsection, we reflect on the development trade-offs and limitations implicit to the approach we followed (6.3).

6.1 Community needs

Community engagement is presented as one central point in each of the two intertwined processes, as it is a phase of validation of the current iteration and definition of the following one. This phase involves presenting the new research findings, tools, and visualizations in various formats as varied as a conference talk, an edit-a-ton workshop or video-call interviews. In the past 2 years, there has been a good deal of dissemination events.[46] In return, it is expected that other community members express their opinions, interests, understandings, and experiences to nurture the Diversity Observatory discourse or tools.

Each of these interactions and collaborations with other community members is essential to collect feedback in the same exact scenarios in which the tools will be used. Requests for features or improvements on the User Interface have been common. Other community members' expectations are usually set on learning and improving the tools, but always in the most transparent and incremental manner - the "wiki-way".[47] Community engagement has both the function of "generative research" and "evaluative research" [43]. Generative to find new opportunities (e.g., requests for inclusion of new diversity categories or new features and visualizations), and evaluative, to validate current elements of the tools and refine the current approach.

This latter is essential at the levels of clarifying the discourse (framing the problems and simplifying the language), prioritizing some analysis (showing the key aspects of the problem), and making the tooling (improving its functionalities and usability). The feedback provided at each interaction is documented and shared across the different active members of the Observatory. The repetition of themes, requests, or concerns and the complexity and associated costs are key to decide the focus of the following iteration.

From the different conversations all over these years, we have identified that other community members needs can be classified into five different groups:

- Understanding the situation and progress for a specific kind of content gap **(situation and progress)**

[46] https://meta.wikimedia.org/wiki/Wikipedia_Diversity_Observatory#Disseminations_timeline
[47] https://meta.wikimedia.org/wiki/The_wiki_way

- Being able to compare different Wikipedia language editions' coverage of some kind of content gap **(comparison)**
- Obtaining discourse and graphical material to explain it to others **(communication)**
- Distinguishing the value in a content gap, i.e., which article should I create **(prioritize)**
- Bridging content gaps more easily and being more efficient at it **(efficiency)**

6.2 Main iterations

The project Wikipedia Diversity Observatory has been developed in three different phases since 2018. We will briefly explain the outcomes presented in each of these iterations, the venues where they were presented and asked for community engagement, and the most common feedback received during these interactions.

6.2.1 First iteration: culture (2018–2019)

In this first iteration, the project was focused on studying Cultural Diversity (its original name was Wikipedia Cultural Diversity Observatory). The goal was to collect the "local content" of each of the more than 300 Wikipedia language editions and create a tool to retrieve lists of relevant articles. In this way we aimed to validate the research results from previous studies [12, 13] with more language editions and raise awareness on the need for increasing cultural diversity, providing some quantitative indicators as it was done for the gender gap.

In the first community interactions, we presented some data tables, simple visualizations, and the first version of the "Top CCC" lists in the regional African communities' conference Wikiindaba[48] (Tunis, Tunisia), the Central and Eastern European communities' conference Wikimedia CEE[49] (Lviv, Ukraine) and Wikimania[50] (Cape Town, South Africa).

6.2.1.1 Website approach validation The idea of dividing the website into "visualizations" and "Tools" was rapidly validated by community members. This way, the first would address the needs of understanding the situation and progress, allowing for comparison and providing evidence, while the second would be focused on organizing action by providing lists of high-priority articles.

6.2.1.2 Top CCC lists usefulness The Top CCC articles lists were embraced by African communities, given that the lists of relevant articles (e.g., Vital articles[51] or List of articles every Wikipedia should have[52]) manually created by Wikipedians in English Wikipedia or Meta-wiki tend to over-represent the Western world.

By having lists of relevant articles according to different criteria but always centred on each cultural context, it was now possible to convince every other language to cover cultural diversity more easily. The discovery was that 127 Wikipedia language editions

[48] https://meta.wikimedia.org/wiki/WikiIndaba_conference_2018
[49] https://meta.wikimedia.org/wiki/Wikimedia_CEE_Meeting_2018
[50] https://wikimania2018.wikimedia.org/wiki/Wikimania
[51] https://en.wikipedia.org/wiki/Wikipedia:Vital_articles
[52] https://meta.wikimedia.org/wiki/List_of_articles_every_Wikipedia_should_have

did not even contain 100 articles geolocated in their corresponding territories or 100 articles on their cultural context.

6.2.1.3 Exporting the lists The Top CCC article lists were used in the annual contest CEE Spring[53] (March, 2019) to select the articles for the 2019 edition. The contest encourages the different languages of the region to create articles about each other's context, including geography, culture, and people. Some Wikipedians pointed out the need for flexibility in providing ways to export the lists to Wikitext or Excel format, so that Wikipedians are able to use them within their daily work.

These were easy upgrades to the tool that were addressed along with some new lists. The Observatory was used to create lists of articles for every country from which participants could choose.[54] Rather than directing participants to the website, they preferred having the lists integrated as tables in the contest page in the Meta-wiki.

6.2.1.4 Creating "Common CCC article lists" One of the participants to the CEE Spring 2019 contest requested to have a list of articles including common aspects of the whole region or at least common to a few countries. This gave place to the tool "Common CCC",[55] which allows searching for articles belonging to the local content of two or more language editions at once.

6.2.1.5 Monitoring changes over time Community members who participated in the CEE Spring 2019 and in Intercultur 2019[56] found the analyses of cultural diversity coverage as something interesting, but observed that without the possibility of monitoring their evolution over time, the metrics were not encouraging progress.

6.2.1.6 Integrating further categories Among all the Wikipedians who used the tools and were inquired to give their opinion about the project, there was consensus on the request for understanding the intersection between gender and cultural contexts, as well as other topics.

6.2.2 Second iteration: gender, geography and minoritized languages (2019–2020)

In the second iteration, the main goal was to address the temporal dimension of the analyses and to expand the data for including gender and geography. We improved the repertoire of visualizations in the Observatory, adding "Diversity Over Time" and dedicated dashboards on gender and geography.

In matters of discourse, we made an important effort to disseminate the value of contributing "local content" and not only cover content from other language editions. In this sense, we presented guidelines[57] giving examples and reasons to create "local content," a chapter on its importance that would be published in the Wikipedia 20 anniversary book.[58]

[53] https://meta.wikimedia.org/wiki/Wikimedia_CEE_Spring
[54] https://meta.wikimedia.org/wiki/Wikimedia_CEE_Spring_2019/Top_CCC_articles
[55] https://wdo.wmcloud.org/common_ccc_articles
[56] Intercultur is similar to CEE, but in this case, it focuses on the Iberian Peninsula and its language communities.
[57] https://meta.wikimedia.org/wiki/Wikipedia_Diversity_Observatory/Guidelines
[58] https://wikipedia20.pubpub.org/pub/26ke5md7/release/15

6.2.2.1 Creating the "Missing CCC dashboard" One of the lessons learnt from the previous iteration was that many language editions do not cover their own local content. This inspired the creation of the Missing CCC articles lists, which provide articles about a language's cultural context that exist in other Wikipedia language editions. These articles lists were very suitable to Indian Wikipedias, which in 2019 organized a contest named "Project Glow"[59] (formerly Project Tiger). This contest originally used lists of topics provided by Google that were generated using the list of local queries in the search engine.[60] However, sometimes the resulting topics could be considered popular (e.g., a smartphone new model) but not necessarily relevant to their context. The Missing CCC articles lists created in 14 Indian languages were provided to the participants to help them identify relevant missing articles that existed in English or other larger Wikipedias.

6.2.2.2 More granularity in time analysis After several discussions at the Wikimania 2019 (Stockholm), it appeared as necessary to be able to have a finer granularity in the analysis than the one provided by the "Diversity Over Time" dashboard. This dashboard was useful to evaluate the impact of the previous Wikimania 2018. However, beyond seeing the distribution of new articles created in the last month, as allowed by the tool, Wikipedians wanted to monitor the edits in the last 24–48 h. This would encourage the creation of the "Recent Changes Diversity" dashboard.

6.2.2.3 Language-based dashboard Another requested improvement was to have the possibility to configure one language-based dashboard for their language, including all the analyses of interest. This appeared as a valuable add-on to the website, after the Diversity Observatory was presented at the regional conference WikiArabia (Marrakech, Morocco).[61] This language-based configurable dashboard would be a valuable step forward in terms of usability, that has not been implemented yet but should be considered.

6.2.2.4 LGBT+ and ethnic groups Some underrepresented groups like LGBT+ and ethnic groups were recurrently in the focus of debates at the Wikimedia Movement Strategy 2030 conversations dedicated to uncovering possible actions and plans to increase diversity[62] in the movement. The degree of organization around these groups is lower than for gender or geography. The requests for including them in the Observatory were aimed at increasing the visibility of these topics and raising awareness on the barriers that prevent their coverage.

6.2.3 Third iteration: LGBT+, ethnic groups and time (2020 - current)

In the third iteration, the main goals were to expand the framework and complete the main topics including LGBT+, Ethnic groups and time, as well as providing new dashboards and features and other improvements to address the needs detected in the previous iteration. The analysis of data on the additional topics would be more

[59] https://meta.wikimedia.org/wiki/Project_GLOW
[60] https://analyticsindiamag.com/google-glow-indian-native-languages-wikipedia/
[61] https://commons.wikimedia.org/wiki/File:The_State_of_Cultural_Diversity_in_Arabic_Wikipedia_2019.pdf
[62] https://meta.wikimedia.org/wiki/Strategy/Wikimedia_movement/2018-20/Working_Groups/Diversity

experimental, as the research on it is scarcer and the level of engagement of the community lower.

6.2.3.1 Ethnic groups data is incomplete We found that data on ethnic groups is incomplete for an extensive and accurate selection of all the content related to them. However, we could still achieve a collection of articles that may serve as a reference point for comparison. Building on this data, we created dashboards like "Ethnic Groups Topic Articles," where biographies or more general topics around each group are listed as a starting point for bridging content gaps across languages. This would be especially useful for events like the "International Roma Day edit-a-thon".[63]

6.2.3.2 LGBT+ content and categorization For the LGBT+ content, we collected all the biographies with a non-heterosexual orientation. Also, we collected all the articles that were categorized as LGBT+ in at least one language edition. For the Wikimedia LGBT+ editors, not only it matters that LGBT+ content exists, but also that it is categorized as such.[64] In fact, only 93 Wikipedia language editions have the "LGBT" related category.[65] We created the dashboard "LGBT+ Articles" which provides a way to retrieve articles related to LGBT+ and sort them according to the number of languages in which they are categorized as such. The feedback received in online events was positive.

6.2.3.3 API request During 2020, community engagement was reduced to online edit-a-thons, and the conversations were mostly focused on different aspects of the use of the tools. We collected some specific requests around their integration in Wikipedia language editions and within other existing tools. For example, enabling third-party applications via data API was requested to update specific metrics in Wikipedia pages using bots. Tools like Fountain tool[66] used in the contest Asian Month[67] to count the number of articles or Bytes added during a contest would also benefit from querying the Diversity Observatory API to be able to see the diversity of articles created. Generally, addressing editors' needs does not only imply making the dashboards more usable, but also providing ways to incorporate metrics and knowledge into other tools and spaces.

6.2.3.4 Diversity in Wikimedia education Identifying content gaps is relevant to those organizations of the movement that are aimed at fostering partnerships or introducing Wikipedia as part of the education system. For example, in the last quarter of 2020 we were contacted by the Wikimedia Foundation education department to create three modules that would be used to explain how to read and edit Wikipedia in the classrooms. We created in the teaching materials a specific section called "Diversity Observatory",[68] encouraging the creation of local content, and translated it to English,

[63] https://meta.wikimedia.org/wiki/International_Roma_Day_Edit-a-thon_2020
[64] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_LGBT_studies
[65] https://en.wikipedia.org/wiki/Category:LGBT
[66] https://meta.wikimedia.org/wiki/Fountain_tool
[67] https://ca.wikipedia.org/wiki/Viquiprojecte:Asian_Month
[68] https://diff.wikimedia.org/2021/04/02/a-three-module-teachers-guide-about-reading-wikipedia-in-the-classroom-is-now-available-on-commons/

Spanish, Arabic, and Tagalog. Along with other texts, the module "allowed teachers to reflect on the importance of cultural representation online, the challenges in accessing sources of information, and building community knowledge responsibly." As of April 2nd, 2021, more than 7000 teachers engaged with the content of the program by accessing the resources and joining live training sessions.

6.3 Development trade-offs

Prior to every new iteration, we acknowledged the different development trade-offs. Most typically, the question is between addressing the new opportunities (generative research) and addressing the improvement of the tooling (evaluative research). In the first iterations, we prioritized fulfilling the requests for expanding the analysis on more topics (e.g., gender, geography, LGBT+, ethnic groups, time, etc.) rather than aiming at giving a polished end-product.

We took note of every usability issue and potential new functionalities. Still, we focused on uncovering the advantages of having one framework to measure the different content gaps, considering that there are some specialized tools for gender. Even though there are always potential new topics, the Diversity Observatory addresses all the topics of interest of the non-geographical Wikimedia user groups[69] (e.g., Wikimedia LGBT+, Gender-related, etc.)

Focusing on developing the framework has been a conscious choice on this particular trade-off between exploring new data and polishing the product. Similarly, when analysing the data in the search for valuable insights, we also preferred exposing many of the visualizations on the website as a matter of openness. The feedback we received allowed us to discard some of them. We realized that the website not only plays a role in providing solutions or insights, but is also an experimental space to invite other community members to reflect on content diversity.

We have noticed that this approach may overwhelm some users, who expect a finished product instead of a research prototype. However, this is a cost for staying open to everyone's feedback, which is important when we are all still learning from the data and what is valuable to Wikipedians. In the future, given that the main topics are already covered and analysed, we expect that it will be possible to reduce the complexity of analyses and metrics to those few that users find particularly valuable.

7 Conclusions

As the Wikimedia movement strives to increase diversity as part of the strategic goals for 2030,[70] the Wikipedia Diversity Observatory is a research project that provides tools and recommendations to bridge content gaps, either by encouraging editors to enrich the representation of their cultural context, by suggesting relevant content from other cultural contexts, or by fostering the creation of biographies about women and minority groups.

We have presented a novel idea leveraging research in the field of Digital Humanities to foster content diversity in peer production. Based on the research on the contextualisation of Wikipedia and its current biases, we built a comprehensive technical

[69] https://meta.wikimedia.org/wiki/Wikimedia_movement_affiliates
[70] https://meta.wikimedia.org/wiki/Strategy/Wikimedia_movement/2018-20/Recommendations/Recommendations

framework to help communities assess content imbalances and take action to reduce them. Language communities differ in their maturity with respect to content diversity, as they have different levels of awareness of the content missing, and of capacity to organize and create it [44]. The Diversity Observatory provides solutions to each of the 304 active Wikipedia language editions, regardless of their community size and current capacity.

One important step in our approach has been continuous community engagement in order to understand the level of involvement and coordination with respect to specific underrepresented categories. While the gender gap receives a lot of attention and has several Wikimedia affiliates in different languages dedicated to it as their primary focus, the culture and the geography gaps tend to be addressed in a more segmented way, such as within contests aimed at covering content about specific territories. We believe that community engagement is essential to work close to the needs and mindset of the editors that are committed to bridge content gaps.

We have shown through several examples how the dashboards available on the project's website can assist editors in their daily work to improve diversity coverage in Wikipedia. In particular, we have illustrated how the visualizations can help to understand and assess the culture gap (4.1), the geography gap over time (4.2) and the gender gap in pageviews (4.3), and how the tools provided can help to identify and bridge the culture gap through lists of relevant articles to be shared across language editions (5.1), lists of relevant articles (5.2) or images (5.3) associated with a culture or territory, and missing in the corresponding language edition.

Furthermore, we have accounted for the iterative process through which the development of the dashboards was driven by input and feedback received from the communities. We have provided a discussion of the main issues and requirements raised by the communities at different iterations, and of the impact of the dashboards on events and contests.

7.1 Future steps

Leaving aside all the improvements uncovered on the dashboards, one area that has not been approached yet by this project is the study of the causes for the lack of diversity in content and its relation to the lack of diversity in contributors. We know little about the different contexts from which editors contribute, and the barriers they have to overcome in order to do so. While this project is focused on creating a cartography for the content, it could benefit from investigating the different barriers and factors that influence contributor diversity. This would help to explain imbalances in both community capacities and content diversity, because the best guarantee that all human knowledge is collected in Wikipedia would be to have a fair and balanced representation of humanity in the movement and its communities.

Abbreviations

CCC: Cultural Context Content; WDO: Wikipedia Diversity Observatory; LGBT+: Lesbian, Gay, Bisexual, Transsexual, and other

Acknowledgements

We thank Andreas Kaltenbrunner, Chris Schilling, Laura Vincze, among others, for their valuable feedback both from an academic and a Wikipedian perspective, and for all their support throughout the development of this project.

Authors' contributions

MM conceived, developed, and disseminated the project. DL supervised the methodology of the study and helped with writing the manuscript. All authors read and approved the final manuscript.

Funding

This work has been partially funded by a project grant from the Wikimedia Foundation for the project "Culture Gap Monthly Monitoring" (March 3rd 2019 – November 10th 2020). The project proposal, along with the feedback and endorsements received from Wikimedia communities, is available online at: https://meta.wikimedia.org/wiki/Grants:Project/WCDO/Culture_Gap_Monthly_Monitoring.

Availability of data and materials

The datasets generated during the current study are available on the project's server: https://wdo.wmcloud.org/databases.
All the code used in the study is available in the WDO GitHub repository: https://github.com/marcmiquel/WDO.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 20 January 2021 Accepted: 1 September 2021
Published online: 01 November 2021

References

1. Bao P, Hecht B, Carton S, Quaderi M, Horn MS, Gergle D. Omnipedia: bridging the Wikipedia language gap: CHI; 2012. p. 1075–84. https://doi.org/10.1145/2207676.2208553.
2. Acey CE, Bouterse S, Ghoshal S, Global AM. Decolonizing the internet by decolonizing ourselves: challenging epistemic injustice through feminist practice: onlineucpressedu; 2021. https://doi.org/10.1525/gp.2021.21268.
3. Wagner C, Graells-Garrido E, Garcia D, Menczer F. Women through the glass ceiling: gender asymmetries in Wikipedia. EPJ Data Sci. 2016;5(1):1–24. https://doi.org/10.1140/epjds/s13688-016-0066-4.
4. Koerner J. Wikipedia has a bias problem. In: Wikipedia @ 20. Cambridge: The MIT Press; 2020. p. 1–11.
5. Yang H-L, Lai C-Y. Motivations of Wikipedia content contributors. Comput Hum Behav. 2010;26(6):1377–83. https://doi.org/10.1016/j.chb.2010.04.011.
6. Jemielniak D, Wilamowski M. Cultural diversity of quality of information on Wikipedias. JASIST. 2017;20(10):247–11. https://doi.org/10.1002/asi.23901.
7. Rizoiu M-A, Xie L, Caetano T, Cebrian M. Evolution of privacy loss in Wikipedia. In: WSDM '16. New York: ACM; 2016. p. 215–24.
8. Gauthier M, Sawchuk K. Not notable enough: feminism and expertise in Wikipedia. 2017;14(4):385–402. https://doi.org/10.1080/14791420.2017.1386321.
9. Roued-Cunliffe H. Forgotten history on Wikipedia. In: Participatory heritage. London: Facet Publishing; 2017.
10. Duncan A. Towards an activist research: is Wikipedia the problem or the solution? 2020. p. 1–14.
11. Bjork-James C. New maps for an inclusive Wikipedia: decolonial scholarship and strategies to counter systemic bias. New Rev Hypermedia Multimed. 2021;10:1–22. https://doi.org/10.1080/13614568.2020.1865463.
12. Miquel-Ribé M, Laniado D. Cultural identities in Wikipedias. New York: ACM; 2016. p. 24–10.
13. Miquel-Ribé M, Laniado D. Wikipedia culture gap: quantifying content imbalances across 40 language editions. Front Phys. 2018;6:234. https://doi.org/10.3389/fphy.2018.00054.
14. Miquel-Ribé M, Laniado D. Wikipedia cultural diversity dataset - a complete cartography for 300 language editions. In: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 13; 2019. pp. 620–9.
15. Hecht B, Gergle D. The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. New York: ACM Request Permissions; 2010. p. 291–300.
16. Graham M, Hogan B, Straumann RK, Medhat A. Uneven geographies of user-generated information: patterns of increasing informational poverty. Ann Assoc Am Geogr. 2014;104(4):746–64. https://doi.org/10.1080/00045608.2014.910087.
17. Karimi F, Bohlin L, Samoilenko A, Rosvall M, Lancichinetti A. Quantifying national information interests using the activity of Wikipedia editors. arXiv. 2015;1503:5522.
18. Samoilenko A, Karimi F, Edler D, Kunegis J, Strohmaier M. Linguistic neighbourhoods: explaining cultural borders on Wikipedia through multilingual co-editing activity. EPJ Data Sci. 2016;5(1):171–21. https://doi.org/10.1140/epjds/s13688-016-0070-8.
19. Warncke-Wang M, Uduwage A, Dong Z, Riedl J. In search of the ur-Wikipedia: universality, similarity, and translation in the Wikipedia inter-language link network. In: OpenSym '12: proceedings of the eighth annual international symposium on Wikis and open collaboration; 2012. p. 20. https://doi.org/10.1145/2462932.2462959.
20. Dittus M, Graham M. Mapping Wikipedia's geolinguistic contours. Digit Cult Soc. 2019;5:147–64. https://doi.org/10.14361/dcs-2019-0109.
21. Sheehan E, Meng C, Tan M, Uzkent B, Jean N, Lobell DB, et al. Predicting economic development using geolocated Wikipedia articles. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining; 2019. p. 2698–706.
22. Hecht B, Gergle D. Measuring self-focus bias in community-maintained knowledge repositories. In: Proceedings of the fourth international conference on communities and technologies; 2009. p. 11–20.
23. Hecht BJ, Gergle D. On the "localness" of user-generated content. In: Proceedings of the 2010 ACM conference on computer supported cooperative work; 2010. p. 229–32.

24. Graham M, Straumann RK, Hogan B. Digital divisions of labor and informational magnetism: mapping participation in Wikipedia. Ann Assoc Am Geogr. 2015;105(6):1158–78. https://doi.org/10.1080/00045608.2015.1072791.
25. Ojanperä S, Graham M, Straumann RK, Zook M. Engagement in the knowledge economy: regional patterns of content creation with a focus on Sub-Saharan Africa. Inf Technol Int Dev. 2017;13:19.
26. Callahan ES, Herring SC. Cultural Bias in Wikipedia content on famous persons. J Assoc Inf Sci Technol. 2011;62(10): 1899–915. https://doi.org/10.1002/asi.21577.
27. Gloor PA, Marcos J, de Boer PM, Fuehres H, Lo W, Nemoto K (2015) Cultural anthropology through the lens of Wikipedia: historical leader networks, gender bias, and news-based sentiment. arXiv preprint arXiv:1508.00055.
28. Apic G, Betts MJ, Russell RB. Content disputes in Wikipedia reflect geopolitical instability. PLoS One. 2011;6(6):e20902. https://doi.org/10.1371/journal.pone.0020902.g001.
29. Ahmed W, Poulter M. Representation of non-Western cultural knowledge on Wikipedia: the case of the visual arts; 2021. https://doi.org/10.20944/preprints202104.0770.v1.
30. Kumar S. A river by any other name: Ganga/Ganges and the postcolonial politics of knowledge on Wikipedia. Inf Commun Soc. 2017;20(6):809–24. https://doi.org/10.1080/1369118X.2017.1293709.
31. Kristiani I. Encouraging indigenous knowledge production for Wikipedia. New Rev Hypermedia Multimed. 2021:1–15. https://doi.org/10.1080/13614568.2021.1888320.
32. Gallert P, Winschiers-Theophilus H, Kapuire GK, Stanley C, Cabrero DG, Shabangu B. Indigenous knowledge for Wikipedia. In: Proceedings of the first African conference on human computer interaction – AfriCHI'16. New York: ACM; 2016. p. 155–9.
33. Hill BM, Shaw A. The Wikipedia gender gap revisited: characterizing survey response bias with propensity score estimation. PLoS One. 2013;8(6):e65782–5. https://doi.org/10.1371/journal.pone.0065782.
34. Reagle J, Rhue L. Gender bias in Wikipedia and Britannica. Int J Commun. 2011;5:21.
35. Konieczny P, Klein M. Gender gap through time and space: a journey through Wikipedia biographies via the Wikidata Human Gender Indicator. New Media Soc. 2018;20(12):4608–33. https://doi.org/10.1177/1461444818779080.
36. Wagner C, Garcia D, Jadidi M, Strohmaier M. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In: Proceedings of the international AAAI conference on web and social media, vol. 9; 2015. p. 1.
37. Graells-Garrido E, Lalmas M, Menczer F. First women, second sex - gender bias in Wikipedia. In: Proceedings of the 26th ACM conference on hypertext & social media; 2015. p. 165–74.
38. Beytía P, Wagner C. Visibility layers: a framework for facing the complexity of the gender gap in Wikipedia content. SocArXiv; 2020. https://doi.org/10.31235/osf.io/5ndkm.
39. Wexelbaum RS, Herzog K, Rasberry L. Queering Wikipedia. 1–20. LGBTQ+ librarianship in the 21st century: emerging directions of advocacy and community engagement in diverse information environments (advances in librarianship), vol. 45. Bingley: Emerald Publishing Limited; 2015. p. 115–39. https://doi.org/10.1108/S0065-283020190000045011.
40. Redi M, Gerlach M, Johnson I, Morgan J, Zia L. A taxonomy of knowledge gaps for Wikimedia projects. arXiv cs.CY:arXiv: 2008.12314; 2020.
41. Science AAFTAO. Promoting an open research culture; 2015. p. 1–5. https://doi.org/10.1126/science.aab3847.
42. Vicente-Saez R, Gustafsson R, Van den Brande L. The dawn of an open exploration era: emergent principles and practices of open science and innovation of university research teams in a digital world. Technol Forecast Soc Change. 2020;156:120037. https://doi.org/10.1016/j.techfore.2020.120037.
43. Goodman E, Kuniavsky M, Moed A. Observing the user experience: a practitioner's guide to user research, second edition; 2012. p. 1–601.
44. Miquel-Ribé M. The sum of human knowledge? Not in one Wikipedia language edition. Wikipedia @ 20. Cambridge: The MIT Press; 2020.

Publisher's Note