The Wikipedia Diversity Observatory

A Project to Identify and Bridge Content Gaps in Wikipedia

Marc Miquel-Ribé
Universitat Pompeu Fabra, Barcelona, Catalonia
marcmiquel@gmail.com

David Laniado
Eurecat, Centre Tecnològic de Catalunya
david.laniado@eurecat.org

ABSTRACT

In this paper we present the Wikipedia Diversity Observatory, a project aimed to increase diversity within Wikipedia language editions. The project includes dashboards with visualizations and tools which show the gaps in terms of concepts not represented or not shared across languages. The dashboards are built on datasets generated for each of the more than 300 language editions, with features that label each article according to different 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.

CCS CONCEPTS

• Human-centered computing → Empirical studies in collaborative and social computing → Human-Computer Interaction (HCI) → Collaborative content creation

KEYWORDS

Diversity, Culture Gap, Gender Gap, Online Collaboration, Wikipedia, Digital Humanities, Data Visualization.

ACM Reference format:

Marc Miquel-Ribé and David Laniado. 2020. The Wikipedia Diversity Observatory: A Project to Identify and Bridge Content Gaps in Wikipedia. In *Proceedings of the International Symposium on Open Collaboration (OpenSym 2020). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/3412569.3412866

OpenSym 2020, August 25–27, 2020, Virtual conference, Spain

ACM ISBN 978-1-4503-8779-8/20/08$15.00
https://doi.org/10.1145/3412569.3412866

1 Introduction

Wikipedia is among the largest information repositories on the Internet that are both multilingual and created through collaborative effort. Its prime objective[1] is to "give free access to the sum of all human knowledge" and, consequently, it exists in as many as 309 languages. Even though the language communities make the projects grow on a constant basis, the content does not represent the existing diversity in peoples, places, and cultures of the world; furthermore, there is a gap between language editions and articles often are not shared, or remain even unique to one language [1]. The creation of articles in Wikipedia language editions is spontaneous and non-directed. Several studies showed that cultural and geographical factors influence the topical distribution of content in Wikipedia language editions [2, 5, 6]. In fact, the most active editions tend to represent extensively the context where the language is spoken, dedicating articles to a variety of topics, but fail to ensure a minimum coverage of the other languages' related cultural and geographical context (the culture gap) [5]. Likewise, there exist other biases like the gender gap [3, 4] resulting in a lower percentage of women in biographies.

In the past years, there has been increasing awareness of the issue of diversity in the communities. The Wikimedia Foundation initiated a Movement Strategy Process and one of the resulting two goals that have been 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". Even though editors maintain their editorial freedom, 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. Beyond these initiatives, tools to monitor different kinds of gaps in Wikipedia content have started to be developed, especially for the gender gap [3]. Other content gaps take longer to be measured because of their complexity. At the same time, no project is aimed at both showing the gaps and

[1] https://en.wikipedia.org/wiki/Wikipedia:Prime_objective

[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/

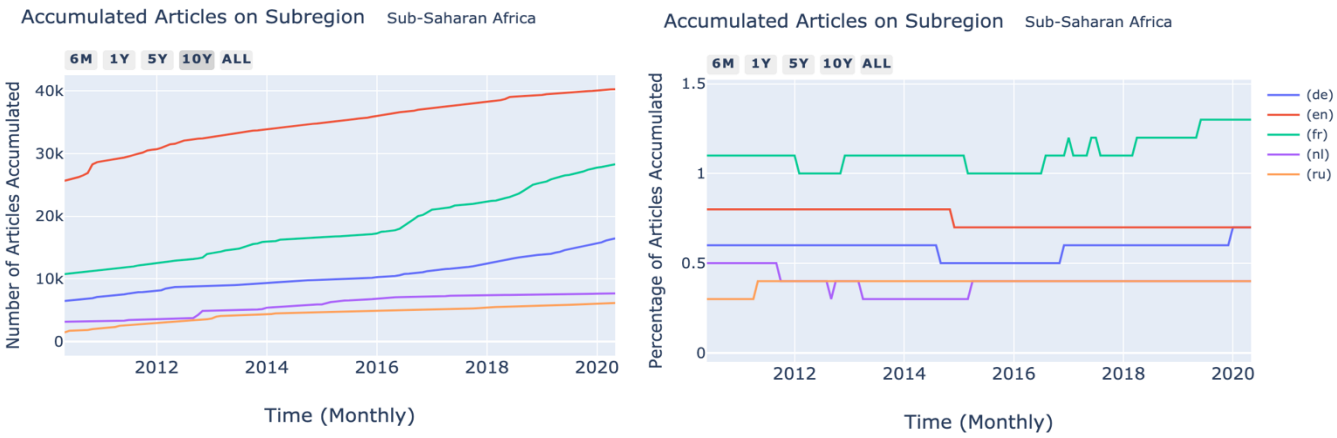

Figure 1: Monthly evolution of the overall number of articles geolocated in Sub-Saharan Africa in the German, English, French, Dutch and Russian Wikipedia, in absolute numbers (left) and normalized by all geolocated articles (right).

providing suggestions to bridge them, so that editors can immediately act.

As a solution to this problem, we present the Wikipedia Diversity Observatory, a project that opens a new space for both scholars and Wikipedia editors to identify and act to bridge the gaps. The project addresses 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 in order to bridge the 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 shows the importance of integrating easy to use dashboards into the community daily activities in order to constantly raise awareness by showing progress and proposing solutions.

2 Approach

The Wikipedia Diversity Observatory[7] approach is three-fold. Firstly, we created a dataset for each Wikipedia language edition 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, the relationship from an article as belonging to the language's related topics requires a more sophisticated method. In this case, 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. Then, we introduce all of them into a machine learning classifier to obtain the final selection of articles belonging to a language context, following the approach described in [6]. The resulting dataset is available in different formats (e.g. Sqlite3 and CSV)[8] [9] and is computed on a regular basis.

Secondly, based on the dataset produced for each Wikipedia language edition, we computed some basic statistics for groups of articles representing content associated to a language, to 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 between them and with larger groups of articles (e.g. an entire language edition, or articles created during the past month). The amount of local content, or content dedicated by each language edition to its associated territories, people and culture (named Cultural Context Content or CCC) was found to represent on average the 25% for the 40 largest Wikipedia language editions [5]. However, for 145 Wikipedia language editions, CCC is below 10%, which clearly points out a problem of representation. In fact, 92 Wikipedia language editions do not even have 100 articles geolocated in their corresponding territories.

Thirdly, we created dashboards with visualizations and tools that use the datasets and statistics generated. 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 geo-located 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

[5] The project visualizations and tools are available at http://wcdo.wmflabs.org/

[6] The project code is available at https://github.com/marcmiquel/wcdo

[7] https://meta.wikimedia.org/wiki/Wikipedia_Diversity_Observatory

[8] https://wcdo.wmflabs.org/databases

[9] https://doi.org/10.6084/m9.figshare.7039514.v3

specific lists of articles and other content suggestions to foster the creation, improvement and exchange of content.

There are dashboards dedicated to different kinds of gaps (Cultural Gap, Geographic Gap, and Gender gap), to article reading measures (Last Month Pageviews), and to temporal analysis of created articles (Diversity Over Time). Other dashboards provide: the most relevant articles from each Wikipedia language edition based on article metrics such as number of page views, edits or editors, and according to a long array of categories relevant for content diversity, including culture, gender or geography (Top CCC Diversity Lists); articles shared between the cultural contexts associated with two language editions (Common CCC); or articles that do not exist in a language edition although they are part of its context, while they exist in larger language editions (Missing CCC). Other tools are aimed at finding the most used images that exist for an article but are missing for a language edition (Visual CCC), the languages in which a list of articles is more complete and thus can help to expand them in other languages (Incomplete CCC) or simply the articles for any specific topic in another language and their relevance features (Search CCC).

2.1 Visualizations

Since the Diversity Observatory database categorizes all the Wikipedia language editions articles according to different types of gaps, we can visualize them both longitudinally and transversely. For example, in the dashboard Diversity Over Time, we can see the creation of articles for one or more diversity categories in multiple language editions. 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 Figure 1 we see the creation of articles geolocated in Sub-Saharan Africa over the past twelve years in five of the largest Wikipedia language editions. While on the left graph we see the growth in 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 language 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 articles creation, as the percentages remain stable.

2.2 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 the dashboards can help bridge the culture gap.

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

N°	Yoruba Article Title	Edits	Editors	Creation Date	Related Languages	Catalan Article Title
1	Genevieve Nnaji	62	12	2009-09-24	es, en, fr, it	Genevieve Nnaji (label)
2	Quincy Olasumbo Ayodele	36	3	2016-07-12	en	Quincy Olasumbo Ayodele (translation)
3	Funmilayo Ransome-Kuti	24	6	2009-12-10	es, en, fr, it	Funmilayo Ransome-Kuti
4	Salawa Abeni	21	6	2011-06-18	es, en	Salawa Abeni (translation)
5	Ngozi Okonjo-Iweala	20	7	2008-10-08	es, en, fr	Ngozi Okonjo-Iweala
6	Agbani Darego	17	6	2009-12-19	es, en, fr, it	Agbani Darego (translation)
7	Oreoluwa Lesi	14	2	2018-10-13	en	
8	Chimamanda Ngozi Adichie	13	10	2009-12-26	es, en, fr, it	Chimamanda Ngozi Adichie
9	Nkechi Justina Nwaogu	13	3	2011-06-18	en	Nkechi Justina Nwaogu (label)
10	Onyeka Onwenu	13	4	2011-06-18	en	Onyeka Onwenu (translation)

Figure 2: Interactive table showing a list of articles on women biographies related to Yoruba culture, sorted by 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.

Case 1: Culture Gap (Top CCC Diversity Lists): sharing the topics related to the language context across language editions. The Top CCC Diversity Lists [10] assist editors in discovering valuable articles from each language's cultural context, and allow them to immediately see their coverage by other language editions. Since the Top CCC Diversity Lists are associated to a language in origin, they address specifically the culture gap in the lack of articles about that language's related topics, but 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 Figure 2 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 rest of columns provide selectable article features such as the number of 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.

[10] https://wcdo.wmflabs.org/top_ccc_articles/

Missing CCC Articles in Wolof Wikipedia on people

Nº	Language	Title	Editors	Pageviews	Interwiki	Bytes	Lang Label
1	en	Tacko Fall	118	53384	5	10.8k	fr Tacko Fall
2	en	Patrice Evra	1774	7346	63	133.0k	fr Patrice Évra
3	en	Idrissa Gueye	212	5512	38	14.2k	fr Idrissa Gueye
4	en	Patrick Vieira	1570	3271	61	83.1k	wo Patrick Vieira
5	en	El Hadji Diouf	1592	2086	36	55.0k	fr El-Hadji Diouf
6	en	Papiss Cissé	960	1400	34	34.1k	fr Papiss Cissé
7	en	Macky Sall	152	1068	48	31.1k	fr Macky Sall
8	en	Mame Biram Diouf	675	732	37	36.1k	fr Mame Biram Diouf
9	en	Dame N'Doye	438	689	27	17.1k	fr Dame N'Doye
10	en	Gorgui Dieng	126	632	21	19.3k	fr Gorgui Dieng

Figure 3: Interactive table showing a list of biographies related to Wolof culture, existing in other language editions and not in Wolof.

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 in the case of covering the diversity in the rest of language editions, we observed that minor language editions do not sufficiently 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. 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. When using the Missing CCC tool[11] to search for articles from the Senegal context that are missing in Wolof Wikipedia, we find this article in the 7th position of the results. This article exists in 48 language editions including English (Figure 3). Possibly, the creation of articles in the Wolof Wikipedia is 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 language editions, sort them by relevance and identify those that may deserve more urgently being created.

3 Conclusions

We have presented a novel idea leveraging research in the field of Digital Humanities to foster content diversity in peer-production. We have provided a comprehensive technical framework to assess content imbalances in Wikipedia. As the Wikimedia movement strives to increase diversity as part of the strategic goals for 2030, 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, or by suggesting relevant content from other cultural contexts. The proposed approach provides solutions to each of the 300 Wikipedia language editions, regardless of their community size and current capacity.

Impact. Feedback provided by the communities has been essential to adjust the tools in order to provide the best-prioritized lists of articles as well as to polish the user interface. Any Wikipedian can suggest new topics for lists of top priority articles to intersect with different categories relevant to overall content diversity. Thanks to constant dissemination in Wikimedia conferences, the use of the tools and visualizations has spread to community initiatives of cross-language article exchange, like the Wikimedia CEE Spring[12] and Intercultur[13]. Beyond community activity, the tools are relevant for initiatives such as the ones aimed at fostering partnerships, or synergies with the education system.

Future Steps. While this project is focused on creating a cartography for the content, it could benefit from investigating the different barriers and factors that influence editing and diversity. This would help to explain imbalances in both community capacities and representation of diversity. Because the best guarantee that all human knowledge is collected in Wikipedia is to have a fair representation of humanity in the movement and its communities.

REFERENCES

[1] Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn and Darren Gergle, 2012. *Omnipedia: bridging the wikipedia language gap*. In Proceedings of CHI '12.

[2] Brent Hecht and Darren Gergle, 2010. *The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context*. In Proceedings of CHI '10.

[3] Maximilian Klein and Piotr Konieczny, 2018. *Gender gap through time and space: A journey through Wikipedia biographies via the Wikidata Human Gender Indicator*. New Media & Society, 20(12), 4608–4633.

[4] Claudia Wagner, Eduardo Graells-Garrido, David Garcia and Filippo Menczer, 2016. Women through the glass ceiling: gender asymmetries in Wikipedia. EPJ Data Science, 5, 1-24.

[5] Marc Miquel-Ribé and David Laniado, 2018. *Wikipedia Culture Gap: Quantifying Content Imbalances Across 40 Language Editions*. Frontiers in Physics, 6, 54.

[6] Marc Miquel-Ribé and David Laniado, 2019. *Wikipedia Cultural Diversity Dataset: A Complete Cartography for 300 Language Editions*. In Proceedings of ICWSM.

[7] Anna Samoilenko, Fariba Karimi, Daniel Edler, Daniel Edler, Jérôme Kunegis and Markus Strohmaier, 2016. *Linguistic neighbourhoods: explaining cultural borders on Wikipedia through multilingual co-editing activity*. EPJ data science 5.1 (2016): 9.

[8] Morten Warncke-Wang, Anuradha Uduwage, Zhenhua Dong, and John Riedl, 2012. *In search of the ur-Wikipedia: universality, similarity, and translation in the Wikipedia inter-language link network*. In Proceedings of OpenSym 2012.

[11] https://wcdo.wmflabs.org/missing_ccc_articles/

[12] https://meta.wikimedia.org/wiki/Wikimedia_CEE_Spring_2020

[13] https://meta.wikimedia.org/wiki/Intercultur/es