Diversity in a Language-Independent Wiki:

Six Design Requirements and Goals to Embed a Diversity Mindset

Marc Miquel-Ribé[1]

[1] Universitat Pompeu Fabra, Barcelona, Catalonia
marcmiquel@gmail.com

Abstract. Abstract Wikipedia is a Wikimedia Foundation project comprised of two repositories, one of functions and another of abstract content. Its goal is to collaboratively create functions that convert content from multiple Wikipedia language editions into an abstract language, allowing editors to maintain it, and feeding it back to each of these language editions so they can expand more rapidly and efficiently. In this paper, we argue that it is necessary to design the functions and user interfaces so that they encourage contributors to actively work on content diversity across Wikipedia language editions. Based on different Wikipedia studies on editing dynamics and content contextualisation, we suggest six design goals and requirements: topic identification, gap bridging, perspective modelling, items representation, article composition, and dispute visualization. Finally, we discuss the importance of helping smaller languages grow and fostering knowledge diversity through the exchange of existing contents.

Keywords: Abstract Wikipedia, Diversity, Content Gaps, Online Collaboration.

1 Introduction

Wikipedia exists in more than 300 language editions, but their distribution of content is unequal. We know there is a gap between language editions as they do not cover each other's content. Even large language editions like English and German only overlap in around a third of their articles. Yet, we also know that there exist dynamics of importing and exporting content from one language to others, aimed at both expanding existing articles and generating new ones.

Based on this, a project named Abstract Wikipedia has been announced[1] in July 2020 to enhance the creation and expansion of articles across Wikipedia language editions. The project proposes a repository of functions that allow capturing content from language editions, converting it to an abstract language, storing it into a language-independent content repository, manipulating it, and finally translating it back to a natural language to be integrated into a Wikipedia language edition [1].

Having a set of functions to operate on multiple Wikipedia language editions and a language-independent repository to store and draw content from seems like a logical solution. A single repository may be less expensive to maintain, and its content can help

[1] https://meta.wikimedia.org/wiki/Abstract_Wikipedia/July_2020_announcement

grow multiple projects at the same time. However, formalising a global and language independent repository could inherently limit the diversity emerging in each language edition, as all the main operations of importing and exporting content would possibly be based on the preferences and biases of its contributors, who might be a less diverse subset of the entire Wikimedia movement.

In this paper, we argue that a project with a multilingual scope like Abstract Wikipedia should take a proactive stance towards diversity and design its repository of functions and its repository of content to best address the existing content gaps and biases in Wikipedia languages (e.g. gender, culture, etc.) at article and in-article level (Section 1). In Section 2, we review the Wikipedia literature and propose six requirements and six goals the system should be planned for, without going further into its technical implementation. In Subsection 2.1, we deepen into article-level and explain why it is necessary to encourage topic identification and gap bridging to export and import articles. In Subsection 2.2, we deepen into in-article level to explain the need for modelling and including, in every article, all the available perspectives or points of view; allowing multiple compositions of the same article with different perspectives; and also raising awareness on disputes in articles before every action. In Section 3, we argue that if these goals were properly addressed (i.e. through some functions and user interfaces), the new repositories could grow faster, fulfilling their unique opportunity to increase the content diversity in Wikipedia language editions.

2 System Design Requirements

In the following two sections, we want to propose a set of six design requirements to ensure that Abstract Wikipedia new repositories are more suited to improve content diversity in all Wikimedia projects at both article and in-article level. Each subsection is entitled according to a system design requirement that is motivated by some research findings and aims at fulfilling a specific design goal.

2.1 Article-Level

More diversity at article-level means that each Wikimedia project, regardless of its size, represents the existing variety of knowledge in Wikimedia. Hence the functions of Abstract Wikipedia should help editors search for groups of articles on a particular topic, to work on them in the language-independent repository, and later, if they wish, to be able to create an article from scratch or improve an existing one in a Wikipedia language edition. In the following Subsection 2.1.1, we will present the two design requirements of topic identification and gap bridging that should enable both to import and export articles from and to Wikipedia language editions.

2.1.1 Topic Identification and Gap Bridging

If we want Wikipedia to gather "the sum of human knowledge"[2], we must inevitably forget about an elitist selection of "universal content" that once motivated the first encyclopaedias. Instead, we need to be able to collect content about every topic. With

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

this comes the challenge of identifying topics and writing articles about them. Even though this is often done through manually generated lists of articles, browsing categories, and navigating Wikidata, it is often challenging to locate a gap, understand its importance and try to bridge it. Abstract Wikipedia as a project proposes to help exchange content across language editions. But what kind of content? The most genuine and unique contribution from each Wikipedia language edition to the global content diversity in Wikimedia is "local content". As it has been stated [2], it is "the group of articles in a Wikipedia language edition that relates to the editors' geographical and cultural context (places, traditions, language, politics, agriculture, biographies, etc.).". Local content takes in average a quarter of the 40 largest Wikipedia language editions (e.g. half of English Wikipedia and 10% of Dutch Wikipedia). It tends to be more edited and developed in terms of Bytes, references, images, among other features [3]. Interestingly, anonymous editors and administrators tend to dedicate a higher proportion of their edits to this kind of content than regular registered editors [3].

What is less known is the fact that the gap between language editions is mainly due to local content [2]. Wikipedia language editions that cover a minimum of others' local content are those which are either very large or simply linguistically or geographically close. Differently from the content gender gap, which exists on most language editions with a proportion of 80-20%, and that can be bridged with the creation of more women biographies for every male biography created, the content culture gap requires content from the local content of every other language edition. [4] proposed a method and tool[3] to select the local content of every language edition and create lists of 100 to 500 articles through algorithms which rank them according to relevance features (i.e. number of editors, number of pageviews, etc.) and specific topics (e.g. books, places, gender, ethnicity, etc.). These "top priority" lists allow editors to have a reference point to start bridging the gaps. Namely, they are centred on a topic and on content local to a language edition and they show the articles' availability in every other language. In a similar vein, the new functions should ease the process of finding articles about a topic and massively create them in other language editions. In other words, a good design goal would be: **"help editors in the process of finding relevant articles about topics they do not know about and import them to a language edition."**

But how about nurturing the other language editions with topics you are familiar with? In many language editions' communities, 50% of the editors who accumulate between 101-1000 edits are multilingual or have contributed at least once to language editions other than their primary language. Such percentage increases to more than 70% for those who accumulated over 1000 edits, and reaches 90% for those who accumulated over 10000 edits [3]. However, the proportion of edits any of these types of editors dedicates to other language editions is always limited to 1-4% of their total contributions. This is also true for some language editions like Catalan or Basque, whose editors also speak another language, as editors generally focus on a primary language. [2] saw that when editors contribute to other language editions, they sometimes translate local content from their primary language to other language editions. In fact, 90 to 100% of the editors with a flag (e.g. sysops) have acted as exporters of their local content, while only 40% of the multilingual registered editors

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

have done so. Most of the multilingual edits are not dedicated to exporting local content but to creating content in general. Unsurprisingly, articles that receive more attention by exporters tend to be focused on the country, the political or historical figures or cities, rather than on an artist or a typical dish. By encouraging the exporting of specific topics (e.g. local content, women biographies, ethnicity, etc.) we can ensure more diversity in every language edition. Taking this into account, a design goal would be: **"help editors in the process of exporting a group of selected articles on a topic they are familiar with into one or more Wikipedia language editions."**

The process of exporting content from large language editions into smaller ones is one of the main promises of Abstract Wikipedia. One would expect that such content would be composed mainly of lists of Vital articles[4] (i.e. "articles that every Wikipedia should have") or articles from a local content whose editors are especially interested in making it available in multiple language editions. However, does it make sense to export local content from a language edition into other language editions which have not yet created their own? [4] shows us that in 145 language editions, local content takes only 5 to 10% of all the articles and that there are relevant gaps in topics as important as central political figures, or places, among others. In fact, 92 language editions contain less than 100 geolocated articles in the territories where the language is spoken. Moreover, the speakers of these languages are more likely to search and read articles on local topics directly in a large language edition, even when they are also available in their own language. This is perhaps due to the acceptance of another language as of higher status.

Since local content receives most of the pageviews, nothing would benefit these smaller language editions more than receiving assistance in the creation of their own local content articles (rather than creating articles on Julius Cesar or international pop-stars). Local content articles can help these language editions reach higher positions in the search engines, hence potentially attracting new contributors. Similarly to this tool[5] proposed by [5], the Abstract Wikipedia new functions should allow searching for articles about a language local content that do not exist in its related Wikipedia language edition, so that local editors can use the content as starting point to create them.

As seen, the motivation to ensure a global audience to an article drives editors to export articles to multiple language editions. However, when the article does not yet exist in any language edition some editors may wonder whether they should first create it in their local language edition or create it directly in the English Wikipedia or, in this case, in the new content repository. While it seems that in the English Wikipedia the article may receive a wider audience geographically speaking, [2] shows that an article belonging to local content tends to receive more pageviews in its related language edition. Most of the editors might prefer the option of starting it locally, and then translate it into English, which tends to be the preferred choice as a secondary language by multilingual editors [5]. What will happen once the Abstract Wikipedia content repository is settled and functioning is uncertain. The available linguistic resources in order to transform abstract content into natural language and back to abstract might probably determine the different paths contributors will take in order to create new content and broadcast it across the Wikimedia projects.

[4] https://en.wikipedia.org/wiki/Wikipedia:Vital_articles

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

2.2 In-Article Level

More diversity at in-article level means that articles contain a wide variety of perspectives. Hence Abstract Wikipedia functions and interfaces should help editors include all the existing points of views in an article and also make more conscious decisions on which perspective is more widely represented in its text. In the following Subsections 2.2.1 to 2.2.4, we will present four design requirements: perspective modelling, items representation, article composition and dispute visualization.

2.2.1 Perspective Modelling

Neutral Point of View (NPOV) is one of Wikipedia's core content policies. It requires an article to be diverse to include "all the significant views that have been published by reliable sources on a topic", but at the same time, it also suggests that these views should be fairly proportionate to the prominence of their sources. In practice, some perspectives like those related to gender or some specific groups are present in some language editions and underrepresented in others. For Abstract Wikipedia to increase in-article diversity in all language editions, it would be necessary that editors recognize the perspectives included in each version of an article. For this, a good design goal is the following: **"help editors to model all the perspectives of an article in all the language editions where it exists."**

But what is a perspective? A perspective is a unit of meaning. It includes one or more statements that complement each other around a specific topic and expands the general topic of the article. The statement is the smallest unit that compounds a perspective. Following Wikidata's definition of statement[6], we understand it as a Subject-Predicate-Object (item, property, value). Sentences usually include more than one statement. In fact, depending on the number of statements, a perspective can take the form of a sentence, a paragraph or even an entire section. You can only increase the usability of content if you are able to accurately model each perspective as a group of statements, which would be a process that could benefit from automated approaches but require manual revision. In fact, the only way to ensure that all the perspectives of all the versions of an article are included in the Abstract Wikipedia content repository, is if they are well-modelled at a statement level. In the process of writing, editors identify perspectives all time and question the validity or suitability of each statement.

With this modularity, it is expected that some recurring perspectives of the articles become standardised as sections or paragraphs, and always receive the same label. For example, some articles would seem to benefit from standardised sections such as "Notable books", "Demographics of a city", "Filmography", among others. While these would be more general perspectives, others could have the length of a sentence and go further into details, e.g. a biography containing one statement which specifies a person's political affiliation or a paragraph with the main ideological affinities. After a while developing perspectives in Abstract Wikipedia, it might lead to more aligned sections of all the articles on the same topic and across language editions. One could say it would eventually work in a similar way to Infoboxes and Wikidata, fact-based and up-to-date. With more perspective modularity, it is also expected that combined

[6] https://www.wikidata.org/wiki/Help:Statements

with topic identification and article management (see Subsection 2.1.2) it will be possible to detect topic-based perspectives from one or more articles, as a prior step to complementing them. For example, given a selected group of biographies of women, we could search for gender-based perspectives in twenty-century historical articles in the Abstract Wikipedia content repository. Using these perspectives, we could update the articles of those language editions which lack them. Many topics but especially those related to gender, geography, colonised nations or ethnic groups could benefit from processes using the modularity that modelling statements and perspectives offers.

2.2.2 Items Representation

When importing content from a language edition into another, we do it with its biases and gaps. In a comparative study of biographies of Americans and Poles, [6] showed that their versions in the Polish and English Wikipedia differed not only at structural level (i.e. number of categories or length) but also in the amount of personal information, mentions to education and nationality, among other aspects. If the English Wikipedia, which is edited by many multilingual editors, does not contain a more complete set of points of view, who could guarantee us that Abstract Wikipedia content repository will? After all, in the demographics of active editors in all Wikimedia projects the English community is the largest, and we might expect that an Abstract Wikipedia article is initially created after the English Wikipedia version. For this reason, we must design the Abstract Wikipedia content repository in such a way that editors are encouraged to make its articles as complete as possible. So, a design goal is the following: **"help editors to include all the perspectives available in all the language editions' versions of an article."** This could be done by showing gaps and differences in the user interface. In some research studies [7, 8] article similarity is computed by counting the number of links two versions of the same article have in common. This idea has also been explored visually by tools[7],[8]. One allows you to pick two or more versions of the same article in different language editions and show the articles in common in a cloud map [7], while the other shows those that are unique to each language in a different colour [9]. Abstract Wikipedia content repository user interface should remind editors which items (articles or Wikidata Qitems) are used and unused in each article's statements and perspectives, so they complete it with the content from all language editions. Similar strategies should be explored to stimulate the completion of the necessary linguistic resources to transform abstract content into the natural languages.

2.2.3 Article Composition

Even if an Abstract Wikipedia article may contain all the perspectives from all the language editions' versions of it, we cannot escape the fact that some perspectives may be more extensively represented than others. This is due to the process through which this content has been created, the number of contributors from each community, as well as their skills and capacity to contribute to Abstract Wikipedia content repository. If an article that is exported to a language edition contains controversial perspectives to the

[7] http://manypedia.com/

[8] https://omnipedia.northwestern.edu/

eyes of the local community, it could be rejected. We know that some articles accumulate many changes on the first lines, such as deletion of substantives, changes in their order, and the alteration of the number of occurrences [10]. Therefore, we should ask if an article of the Abstract Wikipedia content repository, even when it includes all the language's editions perspectives, can satisfy all the target Wikipedia language edition's communities. It seems inevitable that different human groups speaking the same or a different language consider that perspectives on the same topic deserve a different weight. Therefore, having a single arrangement of all the perspectives in an abstract content article to be exported to other language editions may not be the best strategy. For this reason, we propose the design goal: **"allow editors to compose multiple versions of an article to avoid spreading a specific group of perspectives to all language editions."** The modelling of the different perspectives explained in subsection 2.2.2 should facilitate the creation of multiple layouts in which editors, for example, could select which statements they feel more comfortable with. Since Abstract Wikipedia's repository content is not meant to be read like that of any other Wikipedia, it should be flexible to allow editors to choose which perspectives they consider they deserve more representation in the article they are exporting, rather than having a single representation of all perspectives to export.

2.2.4 Dispute Visualization

The reason for allowing multiple arrangements and selections of the statements available in a Wikipedia article is that it can prevent reverts after exporting content to a language edition. But such a strategy comes with a trade-off in terms of efficiency, namely the more potential different compositions are, the more difficult it is to massively create articles. Editors may need to evaluate the different perspectives contained in an article before deciding to export. For this reason, it would be recommended to have a "default" composition. Although it may have an implicit bias, it is nevertheless the most usable solution for editors when there are no relevant disputes. Instead, when dealing with perspectives that have been previously in dispute in the original language editions, the user interface could warn the editor before making an export. Therefore, a good design goal is the following: **"make editors aware of the disagreements in articles' perspectives when importing or exporting content to language editions."** Based on previous research [10], it is expected that articles on scientific discoveries, politics and historical figures are the ones which will require a more thorough examination before exporting. Like in this tool[9], the disputes could be highlighted along with their language edition origin, as this could be useful information for the user to choose whether to omit it or include it before exporting content to a version of the article in another language edition.

3 Conclusions

Abstract Wikipedia repositories of functions and of abstract content will bring efficiency and robustness to Wikimedia, as it guarantees growth through the exchange of contents. Even though the project will be very beneficial to expand these Wikipedia

[9] http://contropedia.net/

language editions whose editors strive to settle a community or even struggle to have access, we need to include them in the decision-making[10]. Likewise, the approach of Abstract Wikipedia to increase diversity is constrained to the existing content in Wikimedia projects, and growing beyond these limits requires more diverse communities. The imbalanced participation in different geographical regions or the enormous gender gap in editors are challenges to be addressed by improving the user experience and providing more safety to editors. Abstract Wikipedia cannot tackle these issues and it should try not to become a technical barrier to existing or new contributors.

We argue that it is necessary to embed diversity in Abstract Wikipedia core functions and content repository interfaces, so that editors are compelled to work on it effortlessly. In this paper we suggested six design goals to enable editors to monitor the gaps, import and export content both at article and in-article level. We proposed six design requirements aimed at fulfilling these goals (topic identification and gap bridging, perspective modelling, items representation, article composition and dispute visualization) based on research studies and practical tools that have been in use in the past decade. We argued that by addressing diversity since the very beginning, the project can become a valuable tool to all those groups of contributors that aim at more knowledge equity, and guarantee an exchange of contents that is both strategic and respectful to the diversity of topics and perspectives.

References

1. Vrandečić, D. (2016). Architecture for a multilingual Wikipedia. arXiv preprint arXiv:2004.04733.
2. Miquel-Ribé, M., Laniado, D. Wikipedia Culture Gap: Quantifying Content Imbalances Across 40 Language Editions. Frontiers in Physics, (2018).
3. Miquel-Ribé, M. Identity-based motivation in digital engagement: the influence of community and cultural identity on participation in Wikipedia (Doctoral dissertation, Universitat Pompeu Fabra), (2017).
4. Miquel-Ribé, M., Laniado, D. Wikipedia Diversity Observatory: A Project to Identify and Bridge Content Gaps in Wikipedia. In Proceedings of OpenSym, (2020).
5. Hale, S. A. Multilinguals and Wikipedia editing (pp. 99–108). WS '14: Proceedings of the 2014 ACM conference on Web science, (2014).
6. Callahan, E. S., Herring, S. C. Cultural bias in Wikipedia content on famous persons. Journal of the American Society for information science and technology, 62(10), 1899-1915, (2011).
7. Massa, P., Scrinzi, F. Manypedia: Comparing language points of view of Wikipedia communities. In Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration (pp. 1-9) (2012).
8. Hecht, B., Gergle, D. The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 291-300), (2010).
9. Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., & Gergle, D. Omnipedia: bridging the Wikipedia language gap. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1075-1084) (2012).
10. Borra, E., Weltevrede, E., Ciuccarelli, P., Kaltenbrunner, A., Laniado, D., Magni, G., ... & Venturini, T. Contropedia-the analysis and visualization of controversies in Wikipedia articles. In OpenSym (pp. 34-1) (2014).

[10] https://en.wikipedia.org/wiki/Nothing_About_Us_Without_Us