A preliminary approach to knowledge integrity risk assessment in Wikipedia projects

Pablo Aragón
Wikimedia Foundation
Barcelona, Spain
paragon@wikimedia.org

Diego Sáez-Trumper
Wikimedia Foundation
Barcelona, Spain
diego@wikimedia.org

ABSTRACT

Wikipedia is one of the main repositories of free knowledge available today, with a central role in the Web ecosystem. For this reason, it can also be a battleground for actors trying to impose specific points of view or even spreading disinformation online. There is a growing need to monitor its "health" but this is not an easy task. Wikipedia exists in over 300 language editions and each project is maintained by a different community, with their own strengths, weaknesses and limitations. In this paper, we introduce a taxonomy of knowledge integrity risks across Wikipedia projects and a first set of indicators to assess internal risks related to community and content issues, as well as external threats such as the geopolitical and media landscape. On top of this taxonomy, we offer a preliminary analysis illustrating how the lack of editors' geographical diversity might represent a knowledge integrity risk. These are the first steps of a research project to build a Wikipedia Knowledge Integrity Risk Observatory.

KEYWORDS

Wikipedia, knowledge integrity, risk assessment

ACM Reference Format:
Pablo Aragón and Diego Sáez-Trumper. 2021. A preliminary approach to knowledge integrity risk assessment in Wikipedia projects. In *The Second International MIS2 Workshop: Misinformation and Misbehavior Mining on the Web (MIS2 workshop at KDD 2021), August 15, 2021, Virtual*. ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

The Web has become the largest repository of knowledge ever known in just three decades. However, we are witnessing in recent years the proliferation of sophisticated strategies that are heavily affecting the reliability and trustworthiness of online information. Web platforms are increasingly encountering misinformation problems caused by deception techniques such as astroturfing [26], harmful bots [2], computational propaganda [24], sockpuppetry [7], data voids [3], etc.

Wikipedia, the world's largest online encyclopedia in which millions of volunteer contributors create and maintain free knowledge, is not free from the aforementioned problems. Disinformation is one of its most relevant challenges [18] and some editors devote a substantial amount of their time in patrolling tasks in order to detect vandalism and make sure that new contributions fulfill community policies and guidelines [1] [13]. Furthermore, *Knowledge Integrity* is one of the strategic programs of Wikimedia Research with the goal of identifying and addressing threats to content on Wikipedia, increasing the capabilities of patrollers, and providing mechanisms for assessing the reliability of sources [27].

Many lessons have been learnt from fighting misinformation in Wikipedia [6] and analyses of recent cases like the 2020 United States presidential election have suggested that the platform was better prepared than major social media outlets [14]. However, there are Wikipedia editions in more than 300 languages, with very different contexts. To provide Wikipedia communities with an actionable monitoring system, this paper introduces a preliminary approach consisting of a taxonomy of knowledge integrity risks in Wikipedia projects, based on a review of the state of the art literature, and an initial set of indicators to be the core of a Wikipedia Knowledge Integrity Risk Observatory.

2 TAXONOMY OF KNOWLEDGE INTEGRITY RISKS

Risks to knowledge integrity in Wikipedia can arise in many and diverse forms. Inspired by a recent work that has proposed a taxonomy of knowledge gaps for Wikimedia projects [16], we have reviewed works by the Wikimedia Foundation, academic researchers and journalists that provided empirical evidence of knowledge integrity risks. Then, we have classified them by developing a hierarchical categorical structure that is presented in Figure 1.

We initially differentiate between **internal** and **external** risks according to their origin. The former correspond to issues specific to the Wikimedia ecosystem while the latter involve activity from other environments, both online and offline. For internal risks, we have identified the following categories focused on either **community** or **content**:

- **Community capacity**: Pool of resources of the community. Risks have been found when, given the size of the corresponding Wikipedia project (volume of articles, edits and editors), there is a shortage of patrolling resources, i.e., editors with elevated user rights (e.g., admins, checkusers, oversighters) [13, 18, 19, 22] and tools (e.g., specialized bots, scripts, MediaWiki extensions, desktop/web apps) [13, 18].
- **Community governance**: Situations and procedures involving decision-making within the community. The reviewed literature has identified risks like the unavailability

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

MIS2'21, August 15, 2021, Virtual

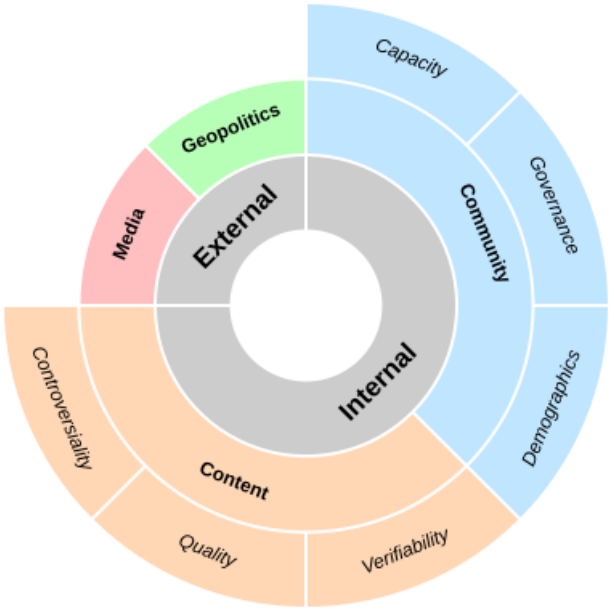

Figure 1: Taxonomy of knowledge integrity risks in Wikipedia.

of local rapid-response noticeboards on smaller wikis [13] or the abuse of blocking practices by admins [17, 19, 22].

- **Community demographics**: Characteristics of community members. Some analyses highlight that the lack of geographical diversity might favor nationalistic biases [17, 19]. Other relevant dimensions are editors' age and activity since misbehavior is often observed in editing patterns of newly created accounts [5, 8, 9] or accounts that have been inactive for a long period to avoid certain patrolling systems or that are no longer monitored and became hacked [13].
- **Content verifiability**: Usage and reliability of sources in articles. This category is directly inspired by one the three principal core content policies of Wikipedia (WP:V) [2] which states that readers and editors must be able to check that information comes from a reliable source. It is referred to in several studies of content integrity [12, 15, 18].
- **Content quality**: Criteria used for the assessment of article quality. Since each Wikipedia language community decides its own standards and grading scale, some works have explored language-agnostic signals of content quality such as the volume of edits and editors or the appearance of specific templates [11, 12]. In fact, there might exist distinctive cultural quality mechanisms as these metrics do not always correlate with featured status of articles [17].
- **Content controversiality**: Disputes between community members due to disagreements about the content of articles. Edit wars are the best known phenomenon that occurs when content becomes controversial [17, 25], sometimes requiring articles to be protected [23].

For external risks, we have identified the following categories:

- **Media** References and visits to the Wikipedia project from other external media on the Internet. Unusual amount of traffic to specific articles coming from social media sites or search engines may be a sign of coordinated vandalism [13].
- **Geopolitics**: Political context of the community and content of the Wikipedia project. Some well resourced interested parties (e.g., corporations, nations) might be interested in externally-coordinated long-term disinformation campaigns in specific projects [13, 21].

3 INITIAL SET OF INDICATORS AND PRELIMINARY ANALYSIS

To capture risks in each category of the proposed taxonomy, we have compiled an initial set of indicators presented in Table 1. The criteria for proposing these indicators are that they should be simple to be easily interpreted by non-technical stakeholders, comparable across wikis, language-agnostic and periodically updatable. For this reason, they are counts of items (e.g., articles, editors, edits, etc.) or distributions of items over informative features.

To illustrate the value of the indicators for knowledge integrity risk assessment in Wikipedia, we provide an example on community demographics, in particular, geographical diversity. Figure 2 shows the entropy value of the distributions of number of edits and views by country of the language editions with over 500K articles. The data has been collected from November 2018 to April 2021. On the one hand, we observe large entropy values for both edits and views in the Arabic, English and Spanish editions, i.e., global communities. On the other hand, other large language editions like the Italian, Indonesian, Polish, Korean or Vietnamese Wikipedia lack that geographical diversity. We should highlight the extraordinarily low entropy of views of the Japanese Wikipedia, which supports one of the main causes attributed to misinformation incidents in this edition [19]. We also notice the misalignment between high edit entropy and low view entropy values in Cebuano and Waray-Waray editions, which might be the result of the large fraction of content produced by bots distributed around the world [22]. It is also remarkable the misalignment of the Egyptian Arabic Wikipedia with much larger entropy values for views than edits.

4 DISCUSSION AND FUTURE WORK

This article has presented a preliminary approach to knowledge integrity risk assessment in Wikipedia projects. We have covered the first steps of an ongoing process with the ultimate goal of building a Wikipedia Knowledge Integrity Risk Observatory. The taxonomy relies on risks detected in previous work on knowledge integrity in Wikipedia, nevertheless, it could be enriched through the review of additional literature on risk assessment in other web platforms.

The analysis shown in this paper has focused exclusively on community demographics. We will extend this work by implementing the rest of the indicators of this and other categories from the taxonomy to assess their informative value. As mentioned above,

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

Table 1: Initial set of indicators for a Wikipedia Knowledge Integrity Risk Observatory.

Risk category	Candidate indicators
Community capacity	Number of articles, editors, active editors, editors with elevated user rights (admins, bureaucrats, checkusers, oversighters, rollbackers); ratio of active editors with elevated user rights; number of specialized patrolling tools; number of AbuseFilter rules.
Community governance	Number of requests in steward's noticeboard; number of global stewards knowledgeable with that language; number of requests for comment (local and meta); ratio of articles for deletion; ratio of blocked accounts (spam, long-term abuse, etc.).
Community demographics	Distribution of views and edits by country; distribution of active editors by age, local activity and cross-wiki activity.
Content verifiability	Distribution of articles by number of citations, number of scientific citations and number of citation and verifiability article maintenance templates, distribution of sources by reliability.
Content quality	Ratio of stub articles; editing depth; distribution of articles by community quality grading, ORES scoring [4], number of editors, number of quality flaw templates, distribution of edits by source type (i.e., editor, newly-registered editor, admin, bot, IP).
Content controversiality	Ratio of locked articles; distribution of articles by controversiality [25], distribution articles by number of comments in discussion page and n-chains in discussion pages [10].
Media	Distribution of mentions/references and visits by online media outlets, social media platforms and search engines.
Geopolitics	Democratic quality scores derived from views and edits by country and well-established country democratic indexes (e.g., [1, 20]).

Figure 2: Entropy values (S) of the distributions of the number of edits and views by country of the Wikipedia language editions, identified by the ISO 639-1 code, with over 500K articles. The graph includes a linear regression model fit.

the current indicators are essentially item counts and distributions of items over features. Future work will also focus on defining advanced metrics while preserving the criteria of ease of interpretation, comparability across wikis and language-agnosticism. Also, indicators should be periodically updatable to allow longitudinal observations. Another future challenge will be to define indicators with finer levels of granularity, that is to say, metrics computed not only on an entire Wikipedia project but on categories, pages, etc.

Last but not least, following the principles of openness, transparency and accountability of Wikimedia [3], we expect to release the Wikipedia Knowledge Integrity Risk Observatory as a dashboard available to the global movement of volunteers. Therefore, we will also focus on designing an open technological infrastructure to provide Wikimedia communities with valuable information on knowledge integrity.

REFERENCES

[1] Coppedge, Michael and Gerring, John and Lindberg, Staffan I and Skaaning, Svend-Erik and Teorell, Jan and Altman, David and Bernhard, Michael and Fish, M Steven and Glynn, Adam and Hicken, Allen and others. 2017. V-dem dataset v7. (2017).

[2] Ferrara, Emilio and Varol, Onur and Davis, Clayton and Menczer, Filippo and Flammini, Alessandro. 2016. The rise of social bots. *Communications of the ACM* 59, 7 (2016), 96–104.

[3] Golebiewski, Michael and Boyd, Danah. 2018. Data voids: Where missing data can easily be exploited. (2018).

[4] Halfaker, Aaron and Geiger, R Stuart. 2020. Ores: Lowering barriers with participatory machine learning in wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–37.

[3] https://meta.wikimedia.org/wiki/Wikimedia_Foundation_Guiding_Principles

[5] Joshi, Nikesh and Spezzano, Francesca and Green, Mayson and Hill, Elijah. 2020. Detecting Undisclosed Paid Editing in Wikipedia. In *Proceedings of The Web Conference 2020*. 2899–2905.

[6] Kelly, Heather. 2021. On its 20th birthday, Wikipedia might be the safest place online. https://www.washingtonpost.com/technology/2021/01/15/wikipedia-20-year-anniversary/

[7] Kumar, Srijan and Cheng, Justin and Leskovec, Jure and Subrahmanian, VS. 2017. An army of me: Sockpuppets in online discussion communities. In *Proceedings of the 26th International Conference on World Wide Web*. 857–866.

[8] Kumar, Srijan and Spezzano, Francesca and Subrahmanian, VS. 2015. Vews: A wikipedia vandal early warning system. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 607–616.

[9] Kumar, Srijan and West, Robert and Leskovec, Jure. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*. 591–602.

[10] Laniado, David and Tasso, Riccardo and Volkovich, Yana and Kaltenbrunner, Andreas. 2011. When the Wikipedians talk: Network and tree structure of Wikipedia discussion pages. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 5.

[11] Lewoniewski, Włodzimierz and Węcel, Krzysztof and Abramowicz, Witold. 2017. Relative quality and popularity evaluation of multilingual Wikipedia articles. In *Informatics*, Vol. 4. Multidisciplinary Digital Publishing Institute, 43.

[12] Lewoniewski, Włodzimierz and Węcel, Krzysztof and Abramowicz, Witold. 2019. Multilingual ranking of Wikipedia articles with quality and popularity assessment in different topics. *Computers* 8, 3 (2019), 60.

[13] Morgan, Jonathan. 2019. Research: Patrolling on Wikipedia. *Report-Meta, Sep* (2019).

[14] Morrison, Sara. 2020. How Wikipedia is preparing for Election Day. https://www.vox.com/recode/2020/11/2/21541880/wikipedia-presidential-election-misinformation-social-media

[15] Redi, Miriam and Fetahu, Besnik and Morgan, Jonathan and Taraborelli, Dario. 2019. Citation needed: A taxonomy and algorithmic assessment of Wikipedia's verifiability. In *The World Wide Web Conference*. 1567–1578.

[16] Redi, Miriam and Gerlach, Martin and Johnson, Isaac and Morgan, Jonathan and Zia, Leila. 2021. A Taxonomy of Knowledge Gaps for Wikimedia Projects (Second Draft). *arXiv preprint arXiv:2008.12314* (2021).

[17] Rogers, Richard and Sendijarevic, Emina and others. 2012. Neutral or National Point of View? A Comparison of Srebrenica articles across Wikipedia's language versions. In *unpublished conference paper, Wikipedia Academy, Berlin, Germany*, Vol. 29.

[18] Saez-Trumper, Diego. 2019. Online disinformation and the role of wikipedia. *arXiv preprint arXiv:1910.12596* (2019).

[19] Sato, Yumiko. 2021. Non-English Editions of Wikipedia Have a Misinformation Problem. https://slate.com/technology/2021/03/japanese-wikipedia-misinformation-non-english-editions.html

[20] Schenkkan, Nate and Repucci, Sarah. 2019. The freedom house survey for 2018: democracy in retreat. *Journal of Democracy* 30, 2 (2019), 100–114.

[21] Shubber, Kadhim. 2021. Russia caught editing Wikipedia entry about MH17. https://www.wired.co.uk/article/russia-edits-mh17-wikipedia-article

[22] Song, Victoria. 2020. A Teen Threw Scots Wiki Into Chaos and It Highlights a Massive Problem With Wikipedia. https://www.gizmodo.com.au/2020/08/a-teen-threw-scots-wiki-into-chaos-and-it-highlights-a-massive-problem-with-wikipedia/

[23] Spezzano, Francesca and Suyehira, Kelsey and Gundala, Laxmi Amulya. 2019. Detecting pages to protect in Wikipedia across multiple languages. *Social Network Analysis and Mining* 9, 1 (2019), 1–16.

[24] Woolley, Samuel C and Howard, Philip N. 2018. *Computational propaganda: political parties, politicians, and political manipulation on social media*. Oxford University Press.

[25] Yasseri, Taha and Spoerri, Anselm and Graham, Mark and Kertész, János. 2014. The most controversial topics in Wikipedia. *Global Wikipedia: International and cross-cultural issues in online collaboration* 25 (2014), 25–48.

[26] Zhang, Jerry and Carpenter, Darrell and Ko, Myung. 2013. Online astroturfing: A theoretical perspective. (2013).

[27] Zia, Leila and Johnson, Isaac and Mansurov, Bahodir and Morgan, Jonathan and Redi, Miriam and Saez-Trumper, Diego and Taraborelli, Dario. 2019. Knowledge Integrity. https://doi.org/10.6084/m9.figshare.7704626