A preliminary approach to knowledge integrity risk assessment in Wikipedia projects

Pablo Aragón

Diego Sáez-Trumper

Internet, August 15, 2021

MIS2 Workshop — KDD 2021

Wikipedia, a reliable source on the Web

How Wikipedia is preparing for Election Day

Wikipedia has a longstanding reputation for inaccuracy. It may no longer be deserved.

By Sara Morrison | Nov 2, 2020, 4:20pm EST

SHARE

Consider clicking on Wikipedia if you're looking for accurate sources of information this Election Day. | Thomas Trutschel/Photothek via Getty Images

[vox.com](vox.com)

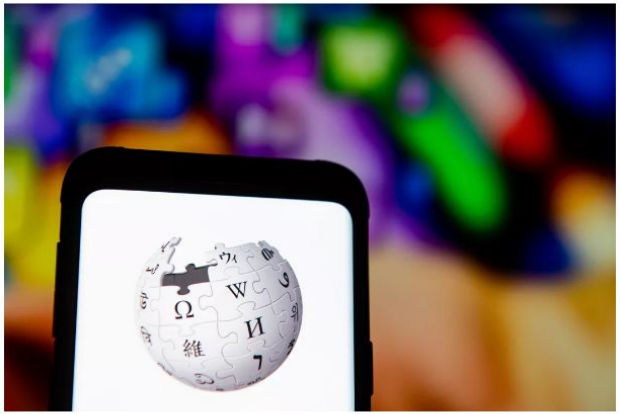
c|net BEST REVIEWS NEWS TECH FINANCE HEALTH HOME CARS DEALS

At 20, Wikipedia has become a refuge from Big Tech's misinformation

The online encyclopedia is an unlikely beacon of reliability.

Richard Nieva Jan. 15, 2021 5:00 a.m. PT

LISTEN · 04:21



Wikipedia celebrates its 20th anniversary.

Mateusz Slodkowski/SOPA Images/LightRocket/Getty Images

cnet.com

Reliability varies across Wikipedias

future tense

Non-English Editions of Wikipedia Have a Misinformation Problem

BY YUMIKO SATO MARCH 19, 2021 • 9:00 AM

slate.com

A Teen Threw Scots Wiki Into Chaos and It Highlights a Massive Problem With Wikipedia

Victoria Song
Published 9 months ago: August 27, 2020 at 4:18 am - Filed to: SCOTS WIKIPEDIA

gizmodo.com.au

Wikimedia Research

Research directions

1. Identify, characterize, and address threats to knowledge integrity

1.1 Research on disinformation campaigns

- Identify projects and topics at risk or particularly vulnerable to coordinated and uncoordinated content manipulation threats

knowledge integrity

Executive summary

The strategic direction of "Knowledge as a Service" envisions a world in which platforms and tools are available to allies and partners to "organize and exchange free, trusted knowledge beyond Wikimedia". Achieving this goal requires not only new infrastructure for representing, curating, linking, and disseminating knowledge, but also efficient and scalable strategies to preserve the reliability and integrity of this knowledge. Technology platforms across the web are looking at Wikipedia as the neutral arbiter of information, but as Wikimedia aspires to extend its scope and scale, the possibility that parties with special interests will manipulate content, or bias to go undetected, becomes material.

In collaboration with multiple partners and collaborators, in 2018-2019 we have started laying foundations for a Knowledge Integrity program through research and development to help our communities represent, curate and understand information provenance in Wikimedia projects more efficiently. We are conducting novel research on why editors source information, and how readers access sources; we are developing algorithms to identify statements in need of sources and gaps in information provenance; we are designing data structures to represent, annotate and analyze source metadata in machine-readable formats as well as tools to monitor in real time changes made to references across the Wikimedia ecosystem. In this white paper, we propose a number of research directions to extend this work over the next 5 years and make progress towards the goals set by the strategic direction.

research.wikimedia.org/knowledge-integrity.html

Research process

Literature review

Review related work on the integrity of knowledge in Wikimedia projects to:

- Identify existing risks
- Group risks into categorical domains

Taxonomy and indicators

Convert risk domains into a taxonomy of risk indicators.

Create a first set of indicators of risks in Wikipedia projects.

Minimum viable product

Create a easy dashboard, based on a sample of the first set of indicators, and validate its informative value with a relevant stakeholder.

Literature review

Related work

Meta: [Research:Wikipedia_Knowledge_Integrity_Risk_Observatory/Literature_review]

WMF reports
Sáez-Trumper (2019); Morgan (2019).

Academic research
Joshi et al. (2020); Spezzano et al. (2019); Lewoniewski et al. (2019); Lewoniewski et al. (2017); Kumar et al. (2016); Kumar et al. (2015); Rogers et al. (2012).

Journalism articles
Sato (2021) on jawiki; Song (2020) on small wikis; Shubber (2014) on ruwiki.

Related work

Morgan, J. (2019). Research: Patrolling on Wikipedia. Research Report.

Risk domain	Excerpt(s)
Community capacity	*Projects with fewer active editors may not be able to ensure real-time review—however, if the volume of edits is correspondingly small, edit review may be a matter of a couple editors performing a daily or weekly 'batch' review of recent changes* *There is no canonical list of all specialized tools that editors have developed and deployed to support patrolling.* (...) *Major bots and assistive editing programs do not work with many projects.* (...) *Smaller wikis tend to have fewer local tool-builders and too-maintainers*
Community governance	*There may not be local rapid-response noticeboards (like AN/I on English Wikipedia) available on smaller wikis*

Related work

Morgan, J. (2019). Research: Patrolling on Wikipedia. Research Report.

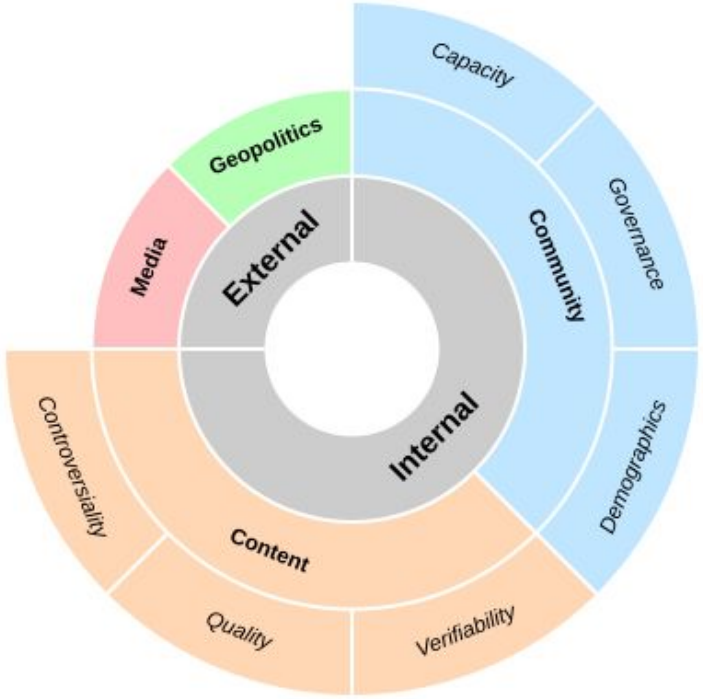
Risk domain	Excerpt(s)
Community demographics	Editors may create accounts and then let them lay inactive for a while (potentially after making a small number of innocuous edits) to avoid certain patrolling mechanisms that call attention to activity by very new accounts and/or accounts with very few edits.
Media	*When an article, or set of related articles, receive a great deal of traffic from social media sites like Facebook or YouTube (which use Wikipedia to fact check controversial UGC) or forums like Reddit (which has been used in the past to coordinate large-scale vandalism), and the article subsequently receives a high volume of edits from IPs or newly registered accounts, this may be a sign of coordinated vandalism*
Geopolitics	*Vandalism can range from persistent disruption-for-disruption's-sake to externally-coordinated long-term disinformation campaigns run by well resourced interested parties such as ideologically-motivated interest groups, corporations, or even potentially nation states.*

Taxonomy and indicators

WIKIMEDIA
FOUNDATION

Taxonomy

Risk Source > Risk Category > Risk Subcategory



Requirements for indicators

- Keep them simple (MVP)
- Easily interpretable
- Comparable across wikis
- Language-agnostic
- Periodically updatable

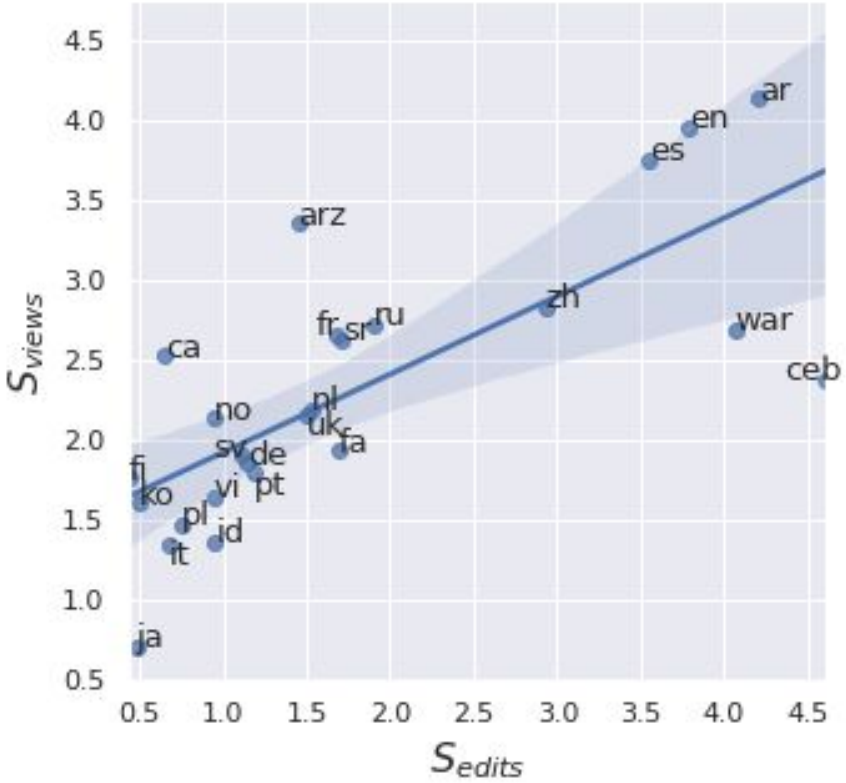
Indicators

Risk category	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]).

Minimum viable product

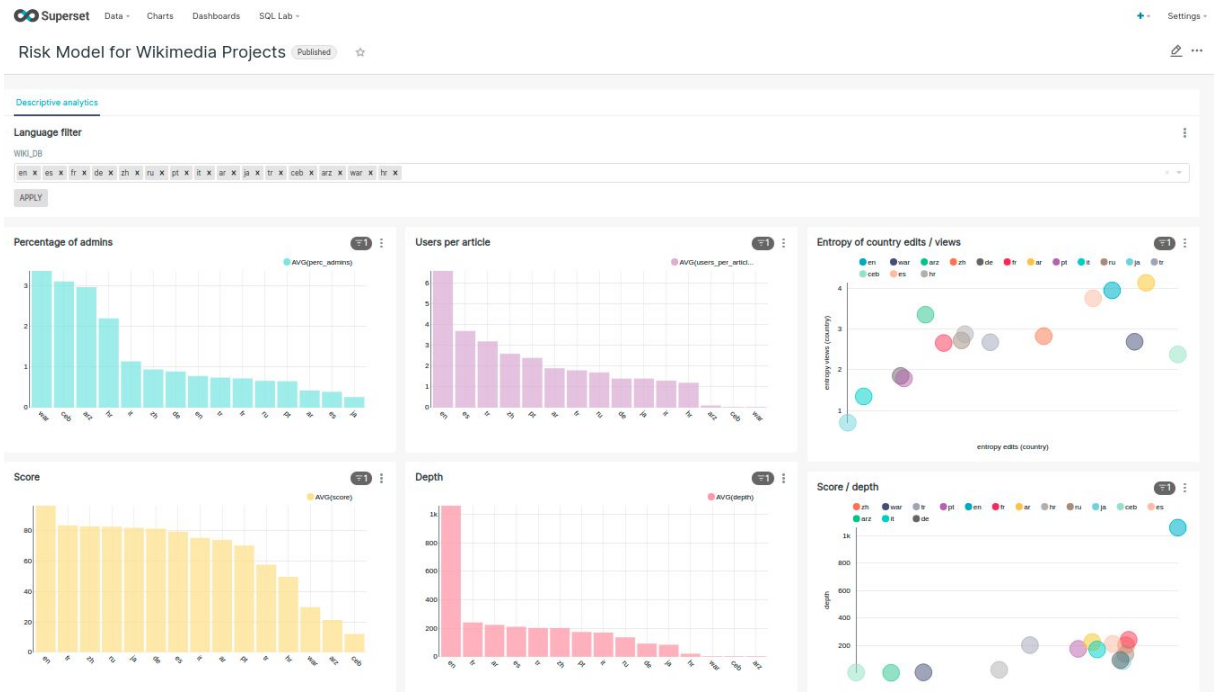
WIKIMEDIA
FOUNDATION

Example: Community demographics



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.

Risk Observatory dashboard



superset.wikimedia.org/superset/dashboard/riskobservatory
(note: *wmf* or *nda* LDAP access is required)

What’s next?
(work in progress)

WIKIMEDIA
FOUNDATION

Limitations and future work

- The risk taxonomy is inspired by works focused exclusively on Wikipedia
  ↳ Review additional literature on risk detection in online platforms

- Many risks generate traces not only in Wikipedia projects
  ↳ Compile data out of Wikipedia ecosystem

- Most metrics are essentially counts and aggregation of distributions
  ↳ Define more advanced and informative metrics while preserving ease of interpretation

- The granularity of metrics has been set to Wikipedia projects
  ↳ Consider how to provide information at more specific levels (e.g. category, page, etc.)

- The current dashboard is only visible to WMF staff / formal collaborators
  ↳ Deploy a technological infrastructure open to the movement in an effective manner

Feedback is a gift

- Literature
- Risk domains
- Indicators
- Datasets
- Technological support
- Stakeholders
- Anything else!

References

- Kumar, S., Spezzano, F., & Subrahmanian, V. S. (2015). Vews: A wikipedia vandal early warning system. In Proceedings of the 21th ACM SIGKDD (pp. 607-616).
- Kumar S, West R, Leskovec J (2016) Disinformation on the web: impact, characteristics, and detection of wikipedia hoaxes. In: Proceedings of the 25th international conference on world wide web, WWW 2016, (pp 591–602).
- Joshi, N., Spezzano, F., Green, M., & Hill, E. (2020). Detecting Undisclosed Paid Editing in Wikipedia. In Proceedings of The Web Conference 2020 (pp. 2899-2905).
- Lewoniewski, W., Węcel, K., & Abramowicz, W. (2019). Multilingual ranking of Wikipedia articles with quality and popularity assessment in different topics. Computers, 8(3), 60.
- Lewoniewski, W., Węcel, K., & Abramowicz, W. (2017). Relative quality and popularity evaluation of multilingual Wikipedia articles. In Informatics (Vol. 4, No. 4, p. 43). Multidisciplinary Digital Publishing Institute
- Morgan, J. (2019). Research: Patrolling on Wikipedia. Research Report. https://meta.wikimedia.org/wiki/Research:Patrolling_on_Wikipedia/Report .
- Rogers, R., & Sendijarevic, E. (2012). Neutral or National Point of View? A Comparison of Srebrenica articles across Wikipedia's language versions. Wikipedia Academy, Berlin, Germany (Vol. 29).
- Saez-Trumper, D. (2019). Online disinformation and the role of Wikipedia. arXiv preprint arXiv:1910.12596.
- Spezzano, F., Suyehira, K., & Gundala, L. A. (2019). DePP: Detecting pages to protect in Wikipedia across multiple languages. Social Network Analysis and Mining, 9(1), 10.

Thanks!

WIKIMEDIA

FOUNDATION