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

At 20, Wikipedia has become a refuge from Big Tech's misinformation
The online encyclopedia is an unlikely beacon of reliability.

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

 vox.com

cnet.com
Reliability varies across Wikipedias

Non-English Editions of Wikipedia Have a Misinformation Problem

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

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

slate.com
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
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 an 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


<table>
<thead>
<tr>
<th><strong>WMF reports</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sáez-Trumper (2019); Morgan (2019).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Academic research</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>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).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Journalism articles</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sato (2021) on jawiki; Song (2020) on small wikis; Shubber (2014) on ruwiki.</td>
</tr>
</tbody>
</table>
## Related work


<table>
<thead>
<tr>
<th>Risk domain</th>
<th>Excerpt(s)</th>
</tr>
</thead>
</table>
| 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 tool-maintainers.  |
| Community governance   | There may not be local rapid-response noticeboards (like AN/I on English Wikipedia) available on smaller wikis.                                                                                                                                                  |
### Related work


<table>
<thead>
<tr>
<th>Risk domain</th>
<th>Excerpt(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community demographics</td>
<td>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.</td>
</tr>
<tr>
<td>Media</td>
<td>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.</td>
</tr>
<tr>
<td>Geopolitics</td>
<td>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.</td>
</tr>
</tbody>
</table>
Taxonomy and indicators
Taxonomy

Risk Source > Risk Category > Risk Subcategory

Requirements for indicators

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

<table>
<thead>
<tr>
<th>Risk category</th>
<th>Indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community capacity</td>
<td>Number of articles, editors, active editors, editors with elevated user rights (admins, bureaucrats, checkusers, overseers, rollbackers); ratio of active editors with elevated user rights; number of specialized patrolling tools; number of AbuseFilter rules.</td>
</tr>
<tr>
<td>Community governance</td>
<td>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.).</td>
</tr>
<tr>
<td>Community demographics</td>
<td>Distribution of views and edits by country; distribution of active editors by age, local activity and cross-wiki activity.</td>
</tr>
<tr>
<td>Content verifiability</td>
<td>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.</td>
</tr>
<tr>
<td>Content quality</td>
<td>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).</td>
</tr>
<tr>
<td>Content controversiality</td>
<td>Ratio of locked articles; distribution of articles by controversy [25], distribution articles by number of comments in discussion page and n-chains in discussion pages [10].</td>
</tr>
<tr>
<td>Media</td>
<td>Distribution of mentions/references and visits by online media outlets, social media platforms and search engines.</td>
</tr>
<tr>
<td>Geopolitics</td>
<td>Democratic quality scores derived from views and edits by country and well-established country democratic indexes (e.g., [1, 20]).</td>
</tr>
</tbody>
</table>
Minimum viable product
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)
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!

@elaragon
References


Thanks!