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.

[Version 1.1 • February 13, 2019]

Table of contents

Cite this document as: Leila Zia, Isaac Johnson, Bahodir Mansurov, Jonathan Morgan, Miriam Redi, Diego Saez-Trumper, and Dario Taraborelli. 2019. *Knowledge Integrity - Wikimedia Research 2030*. doi.org/10.6084/m9.figshare.7704626 [CC BY 4.0]

Desired outcomes

By nature, Wikimedia projects allow anyone sharing our vision to participate in our efforts to share in the sum of all knowledge. This open contribution model has enabled Wikipedia in particular to grow to the size it is now, but it has also allowed parties with special interests to introduce bias and misinformation. The **first outcome** of this program is a set of methods and tools to identify and address coordinated and uncoordinated threats to the integrity of knowledge represented in Wikimedia projects, and empower editors to efficiently monitor the application of content policies at scale.

While Wikipedia's sourcing policies prevent information lacking reliable secondary sources from being considered, these policies have also prevented vast and important domains of knowledge from entering the project. This is particularly critical for cultures that rely on means of knowledge transmission and representation such as oral sources. The **second outcome** of this program is the ability for our communities and movement partners to collaboratively vet and verify knowledge that does not rely on the type of secondary sources currently enshrined in Wikimedia policies. Once we can identify ways of identifying bias or quality issues in such knowledge, we can empower volunteer curators and affiliates to include such knowledge in Wikimedia projects while also helping restrictive content policies to evolve over time.

Research directions

In the next years, we would like to focus the attention of research within the Foundation and among our community partners and academic collaborators, on the following research questions related to Knowledge Integrity:

- How can Wikimedia projects expand the scope and depth of the knowledge domains they serve while retaining their *trustworthiness*?
- How can the Wikimedia movement connect and federate disparate knowledge repositories, subject to different authorities, while ensuring this knowledge remains verifiable in accordance to its sourcing standards?
- How can Wikimedia projects encompass areas of knowledge whose provenance falls outside the current definition of "reliable, secondary sources" as enshrined in their policies?
- And finally, how can the Wikimedia movement scale its curation processes so as to preserve the integrity of its contents in the face of concerted attacks?

We propose to tackle these questions through the following directions:

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

Wikimedia communities have developed over the years robust socio-technical strategies for defining, identifying, and addressing threats to the integrity (e.g. neutrality, verifiability, overall 'quality') of knowledge they create and curate. These communities have policies, processes, as well as tools for dealing with issues of vandalism, non-neutral language, conflicts

of interest, promotional editing, sockpuppets, and spam.

However, traditionally these issues have primarily been identified and addressed at the level of individual actors working alone. In the past decade, the proliferation and widespread adoption of social media technologies have made it easier for people to coordinate their actions in a distributed fashion across multiple platforms. This has paved the way to coordinated campaigns to infiltrate, disrupt, and co-opt movements, communities, and platforms. Wikimedia's mechanisms for addressing knowledge integrity threats were designed to deal with uncoordinated threats in an ad hoc fashion; we don't know how effective they are against—for instance—a targeted campaign to insert disinformation or biased contents into our projects, or to infiltrate leadership roles within a smaller community with the goal of changing community policies or content to reflect the stance or serve the purposes of a corporation, cultural group, or nation state.

Other platforms like Facebook, Twitter, and YouTube have fallen victim to these coordinated efforts. Wikipedia is organized differently from these platforms and may be more robust against coordinated manipulation, but that's no guarantee that it will be effective at combating full-scale "information warfare" from well-funded and more sophisticated disinformation campaigns. There is evidence that groups have proposed and conducted such campaigns in the past. Other projects such as Wikidata, that lack advanced quality control processes, may become particularly vulnerable to these threats.

The proposed objectives for this research direction are the following:

1.1 Research on disinformation campaigns

- Identify projects and topics at risk or particularly vulnerable to coordinated and uncoordinated content manipulation threats;
- Classify qualitatively and quantitatively the characteristics of coordinated editing campaigns, and understand their likely triggers and motivations.

1.2 Detection and timely reporting of potential disinformation campaigns

- Research scalable solutions for real-time monitoring of biased/damaging information being added to Wikimedia contents as a result of coordinated campaigns.
- Develop resources (policies, feedback channels, models, APIs/applications) for community members, WMF staff, and affiliates to identify, timely report and escalate alleged disinformation campaigns.

2. Help contributors monitor violations of core content policies and assess information reliability and bias both granularly and at scale

Despite the effectiveness of Wikimedia's decentralized, collaborative processes for vetting information, the volume of content requiring curation increases at a faster rate than the growth of our projects' contributor base. As a consequence of this, not all components of knowledge represented in Wikimedia projects (sections, citations, sentences, images, categories, factual statements that represent structured data or relations), can be efficiently assessed against content policies. Platforms like ORES provide tools to significantly scale up the ability for curators to monitor a variety of content quality issues in real time.

However, the lack of more granular tools to monitor individual components of Wikimedia contents may threaten the long-term credibility of the project, make its contents vulnerable to misinformation, bias, and overall quality degradation.

We aim to explore the extent to which algorithms can help contributors identify, monitor and fix violations of core content policies at the level of individual information subcomponents. The proposed objectives for this research direction are the following:

2.1 A machine-readable representation of knowledge that exists within Wikimedia projects along with its provenance

- Develop a machine-readable, semantic representation for knowledge available in Wikimedia projects, along with its provenance.
- Design strategies to map existing contents across languages and topics to this semantic representation
- Develop methods to score the confidence of individual knowledge statements, based on available sources in Wikimedia projects and knowledge bases federated with Wikimedia via Wikidata.

2.2 Models to assess the quality of information provenance

- Develop strategies to enrich metadata about sources with machine-readable credibility and reputation indicators, sourced to external authorities, leveraging existing WikiCite work.
- Explore the possibility of using predictive models to help contributors determine whether sources used to back individual knowledge statements are reliable, based on community standards, or if they violate any credibility and reputation indicators, and to facilitate their review at scale.
- Design systems to recommend high-quality sources as potential references or provenance for unsourced (or poorly sourced) statements.

2.3 Models to assess content neutrality and bias

- Identify areas of content violating policies of neutrality and no-original-research.
- Characterize how editors monitor these core content policies, and explore the possibility of training algorithms to score individual knowledge statements to the extent to which they comply or violate these norms.
- Apply these strategies to flagging topic or language areas that may suffer from bias, as a way of recruiting experts to audit and address these content quality issues.

2.4 Research on bias propagation through reuse

- Develop methods to identify and track when biased or inaccurate contents from Wikimedia projects are reused and propagated by third parties.
- Design and test frameworks to escalate the review and potential correction of these contents in Wikimedia projects, in collaboration with downstream content consumers and reusers.

3. Represent and verify knowledge in the absence of secondary or printed sources

Wikipedia policies mandate that information that only appears in a single source, or only in primary sources should not be considered. We propose to research and identify solutions to the problem of verifying knowledge in the absence of “secondary sources” as prescribed by current core content policies. This effort would especially benefit smaller Wikipedia language editions or Wikimedia projects that get information from cultural contexts that draw on limited sources or sources that are incompatible with those policies. A

similar problem affects knowledge that cannot be traced to written sources. In many non-Western cultures, the sum of all human knowledge considerably outweighs the sum of printed knowledge. In order to expand Wikimedia contents in these cultural contexts, we submit that contributors may adopt new methods to collect, represent and vet knowledge that falls outside of the scope of current content policies. Such techniques may include oral citations, images, and video records, coupled with structured metadata.

3.1 Research methods to verify knowledge that lacks secondary sources defined by core content policies

- Understand needs and opportunities for groups and communities affected by the lack of sources that meet the requirements of current verifiability policies
- Design and evaluate workflows that combine structured data and rich media repositories to represent non-textual sources, to address the above needs.
- Study how the above structures representing new knowledge sources can be integrated into processes and policies that communities currently use to reach consensus on, and verify individual knowledge statements.
- Conduct research to differentiate legitimate sourcing efforts by communities that rely on non-textual knowledge sources from attempts to create and spread disinformation, hoaxes, or conspiracy theories through Wikimedia projects in areas and topics where secondary sources are available.

4. Characterize the integrity of multimedia knowledge

While issues of integrity and verifiability in the context of the Wikimedia movement have primarily been considered in the context of textual knowledge, people increasingly share and consume information using non-textual media, in the form of images, videos, audio recordings. As Wikimedia projects rely on multimedia support to represent and disseminate knowledge, it is imperative that we monitor multimedia knowledge against the criteria of integrity and reliability discussed in this program.

4.1. Help contributors identify violations of image quality and integrity policies

- Develop a qualitative and quantitative taxonomy of known and anticipated integrity issues affecting non-textual contents.
- Design methods and prototype tools to help perform quality checks on images against policies. Adequately trained algorithmic tools can help contributors detect and flag low-quality images, images that are of poor relevance or introduce cultural biases that are inappropriate in those contexts in which they are used.
- Develop techniques to identify fake or manipulated images/videos. Multimedia forensics tools can be used here to evaluate authenticity of images and videos.