knowledge gaps

Executive summary

In 2030, the world's population is projected to be 8.6 billion, almost 80% of which will live in Africa and Asia. Latin America's population will continue to grow rapidly while population growth in Europe and Northern America—today's largest sources of contributors and readership to Wikimedia projects—will plateau. How can we help Wikimedia projects thrive in a world that is becoming increasingly different from the one we are building for today, both in terms of production and consumption of content?

The Wikimedia movement has identified as a strategic goal supporting "the knowledge and communities that have been left out by structures of power and privilege". In order to meet this goal, we need to understand how to serve audiences, groups, and cultures that today are underrepresented in Wikipedia, Wikidata, Commons and other Wikimedia projects—in terms of participation, access, representation, and coverage.

In 2018-2019, we have begun to advance knowledge equity with a research program to address knowledge gaps. This program aims to deliver citable, peer-reviewed knowledge and new technology in order to generate baseline data on the diversity of the Wikimedia contributor population, understand reader needs across languages, remove barriers for contribution by underrepresented groups, and help contributors identify and expand missing content across languages and topics. In this white paper, we propose research directions that expand this work over a longer time horizon.

[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 Gaps – Wikimedia Research 2030*. doi.org/10.6084/m9.figshare.7698245 [CC BY 4.0]

Desired outcomes

Addressing inequality rooted in power and privilege structures requires interventions that are both programmatic and technological in nature. The aim of this program is to deliver research which—if successful—will provide decision makers and contributors in the Wikimedia Movement with three types of resources to better target and coordinate their efforts.

The first expected outcome of this program is the availability of **qualitative and quantitative knowledge** to understand the nature, scope and impact of knowledge gaps.

The second outcome is a set of **conceptual and analytical tools** allowing multiple players in the Wikimedia movement to make data-informed decisions on how to select, prioritize and assess different types of interventions.

The third outcome is the availability of research to inform the **design and choice of technological and programmatic levers** that can be used to address these priorities.

Research directions

1. Identify Wikimedia knowledge gaps

The primary focus of initiatives aiming to bridge knowledge gaps in the Wikimedia movement has traditionally been on one particular definition of gaps: *whether a piece of content in a specific form exists on a given project or not.* While this narrow definition has been an effective framing for identifying and addressing specific types of content coverage gaps, it falls short on a few important levels.

First, knowledge gaps cannot be fully defined without characterizing and understanding the need and priorities of Wikimedia's direct readership. Today we know, for example, that Wikipedia readers have needs, motivations, and prior knowledge that vary significantly depending on the language edition they read and the human development status of the country they live in. However, focusing on the needs of readers and connecting these needs to the availability of content is mostly absent from current discussions (see for example Warncke-Wang et al, 2015 for early research on this topic).

Second, in projects that depend on user-generated content, content gaps must be understood in the context of the demographics of contributors who produce content. The lack of diversity in certain dimensions (such as race and gender) has been associated with over- or under-representation of certain types of content, and more broadly to some of the biases reflected in Wikimedia projects.

Third, any technological or programmatic solution to address these gaps must involve community leaders, experts and community-led initiatives. However, little is known about the program and event organizers, governance experts, policy contributors that currently play critical roles support growth and diversity in a project aspiring to represent the sum of *all* knowledge.

It is clear to us that a better grasp of the different dimensions of knowledge gaps and the possible levers to address them is a required first step.

1.1 A taxonomy of Wikimedia knowledge gaps

- Build a taxonomy of Wikimedia *content* gaps: whether the content is present or not (*selection*), how much coverage it has (*extent*), and whose priorities and perspectives are reflected in the content (*framing*).
- Build a taxonomy of gaps in Wikimedia *readership*. This taxonomy will allow us understand what types of readers we are not engaging with: what are these reader needs and how can we support them?
- Build a taxonomy of Wikimedia *usage* gaps. This taxonomy will allow us to understand gaps in the accessibility of content and knowledge that already exists on the projects.
- Build a taxonomy of Wikimedia *contributorship* gaps: what are the characteristics and focus of our current contributors, and how can we encourage diversity in our contributor base and new forms of contribution–beyond editing?
- Build a taxonomy of primary causes of knowledge gaps, including policies, contributor motivations, digital literacy, cultural differences, and access to sources.

2. Measure and prioritize knowledge gaps

Once we have identified and characterized different types of knowledge gaps, we can measure and prioritize them. When approaching this task, we should consider two caveats: First, knowledge is, by definition, always expanding and, as a result, the effort to quantify knowledge gaps requires constant monitoring. Effective prioritization will require us to be able to measure the scope and impact of different types of knowledge gaps at any point in time. Second, knowledge on Wikimedia projects is multimodal. We must therefore focus on developing research methods that are effective at measuring gaps across different platforms, schemas, and media types. This research program will span the following directions :

2.1 Content gaps inferred from knowledge repositories

- Estimate different types of content (text, images, structured data, etc.) missing from Wikimedia projects by taking into account knowledge that exists across Wikimedia projects, knowledge represented in external repositories (knowledge bases, catalogs, authority files, external sources and reference works), as well as knowledge that has not been documented.

2.2 Content gaps inferred from knowledge audiences

- Provide a map of missing content across Wikimedia projects by taking into account reader needs and motivations, their geographical and cultural contexts, their access to technology and other sources of knowledge.
- Provide a map of missing content across projects taking into account the characteristics, demographics and domain expertise of contributors that are needed to help bridge such gaps.
- Develop a working model of knowledge equity, based on Movement values and goals, that we can use to decide which knowledge gaps to focus on first. Defining what Wikimedia means by

"knowledge equity"can help us avoid introducing our own biases, reinforcing existing biases, or causing other unintended consequences for the population or objective we are trying to protect.

3. Bridge knowledge gaps

The third direction of this program aims to identify programmatic and technological levers the Wikimedia movement can rely upon to address known gaps.

3.1 Programmatic levers

- Conduct research to identify current community practices and feedback mechanisms that can help communities address knowledge gaps efficiently.

3.2 Technological levers

- Design and test technology (such as recommender systems and machine classifiers) to assist contributors in identifying and filling knowledge gaps.
- Research, design and test search, navigation and presentation experiences that assist readers in finding the information they are looking for.
- Research and build reader experiences that adapt as a function of reader needs and literacy.

4. Multimedia knowledge gaps

Internet users are increasingly turning to rich media knowledge sources, e.g. images, audio, and video. Wikimedia's multimedia content is subject to the same biases, gaps, and limitations as our textual knowledge and it is imperative that we address these issues to meet further the Wikimedia Strategy goal of Knowledge Equity. The fourth direction of this program addresses the problem of knowledge gaps for non-textual content. By adopting computer vision technology, we can improve the discoverability of multimedia content across languages, and support the visual enrichment of Wikimedia projects. Platforms like ORES allow Wikimedia contributors and developers to characterize the quality of article text and structured data. To extend this functionality beyond text, we first need to build tools that automatically categorize multimedia content.

4.1 A framework for multimedia classification

- Design and implement a framework that allows researchers and practitioners at the WMF to train image/video classifiers for different use-cases. These classifiers, given an image, produce labels characterizing the content and quality of the image, using both supervised and unsupervised approaches, for example identifying what the image depicts.
- Make classifiers multilingual by linking classifiers' labels to corresponding Wikidata items.
- Design an evaluation protocol to make sure that visual classifiers are accurate, unbiased and inclusive.

4.2 Measuring visual knowledge gaps

- Conduct research to understand the role of media in learning and the current gaps in multimedia contents across projects that learners expect.
- Estimate the proportion of visual knowledge missing across Wikimedia projects.

4.3 Visual enrichment of Wikimedia projects

- Research and design visual search tools to improve discoverability of free-licensed multimedia content.
- Research, design, and test technologies to assist contributors visually enriching Wikimedia projects (e.g. recommend images for Wikipedia articles, or support collaborative video editing).

5. A knowledge equity index

The last research direction aims to create a knowledge equity index to provide detailed metrics about the performance of Wikimedia projects towards the strategic goals of knowledge equity.

Socio-economic indices have been adopted by many organizations, advocacy groups and policy makers to track the effectiveness of specific interventions and to measure the health and progress of the audiences they serve. Such an index can be an essential tool to allow Wikimedia communities to measure the performance of their activities against the strategic direction. It can also help inform decisions on policy or effort allocation across the movement. Knowledge gaps are an important component of this index, along with metrics related to the “health” of each project.