foundations

Executive summary

Wikimedia projects are created and maintained by a vast network of individual contributors and organizations with different roles and expertise. The Wikimedia Foundation, including Wikimedia Research, plays an important role in supporting these efforts, but our internal capacity and expertise will always be more limited than those of the Movement as a whole. Tackling the strategic challenges ahead requires an investment in *foundational social and technical infrastructure* that individuals, groups, and organizations across the Movement can use. In this paper, we identify several key *capacity gaps* that impede our shared ability to focus research efforts towards addressing Knowledge Equity and Knowledge as a Service effectively and at scale.

We see an urgent need for increasing the development and dissemination of foundational resources to grow research capacities across the Movement. These foundational resources take many forms: new tools for developing scientific knowledge about projects and contributors; new open data resources and improved tools for working with them; new methods and guidance for mission-aligned research and technology development; and outreach activities designed to foster a healthy, diverse, and dynamic community of researchers to be part of the Wikimedia Movement.

[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. *Foundations - Wikimedia Research 2030*. doi.org/10.6084/m9.figshare.7704629 [CC BY 4.0]

Desired outcomes

With more than 1.5 billion unique devices visiting Wikimedia projects every month and more than 600 thousand monthly registered editors (and many more unregistered) contributing from across the globe and in many different languages, these projects form an immense ecosystem comprising many actors: editors and content providers, readers, movement organizers and affiliate organizations, researchers and academic institutions, developers, cultural organizations and educational institutions, third-party data users, governance and policy stakeholders, and beyond.

While this diversity is essential to Wikimedia's vision of a world in which every human being can share in the sum of all knowledge, that same diversity of stakeholders can create challenges in understanding and adequately serving their needs.

How do we empower the different stakeholders in the Wikimedia Movement to conduct their work more effectively and how can we scale such efforts and go beyond specific country or language communities. In this paper, we discuss four desired outcomes that foundational investments in research outreach and infrastructure can support.

The first outcome is to empower researchers, developers, and decision makers with the ability to collect timely and granular information from the particular groups and communities they serve.

The second outcome is an effective democratization of data and research resources in the Wikimedia ecosystem, making them accessible to everyone, particularly in communities that lack dedicated research or analytical capacity.

The third outcome is a set of principles for the ethical adoption of machine learning and artificial intelligence technologies as well as responsible research practices, to ensure that tools and new technology released on Wikimedia platforms do not propagate bias or create disparate impacts against underrepresented groups.

The last outcome is the availability of a set of resources to help decentralize and increase the research capacity and outreach among affiliates and organizations across the Wikimedia movement.

Research directions

1. Collect timely, granular information from particular groups or communities

We lack effective infrastructure for eliciting targeted information from readers and contributors in an efficient manner, and this limits the kind of research, programs, and policies we can implement. We should invest in tools and develop alternative ways of meeting our needs for rapid, scalable community/sub-group input. This will help us iterate on research questions more rapidly and propose solutions for unmet needs of our communities in a timely manner.

The proposed objectives for this research direction are the following:

1.1 More usable and scalable tools to reach various groups of stakeholders and users

- Improve tools for contextual surveys (such as *QuickSurveys*) with enhanced ways to sample users, articles, topics etc.
- Improve tools for labeled data collection (such as *WikiLabels*) to engage different communities in producing and auditing labeled data for machine learning applications.
- Improve recruitment and notification tools (such as *SendBulkEmail*) to contact specific subgroups of contributors identified by activity levels, edit count, registration date, or other metrics.

1.2 Streamlined processes for gathering input and requests from communities

- Create an end-to-end process and system for gathering input from different communities by taking different facets of research, such as privacy, legal, analytical, and methodological factors into account.
- Create workflows for contacting multiple communities at once and with ease. Contacting different communities (with tools like *MassMessage*) requires many iterations and knowledge of different languages. The proposed workflow will allow us to use such tools and gather input more rapidly.

2. Make existing and new research and data resources more accessible and useful for our stakeholders

The Wikimedia Foundation's Research Team routinely produces and maintains (and sometimes co-maintains with other teams) a large number of research resources, tools and datasets to support the creation of new research and insights on Wikimedia projects. Community members, developers and academic researchers also contribute similar tools and resources. Most of these resources are publicly available, but they may not be provided in a way that makes it easy for stakeholders (such as product teams within the Wikimedia Foundation, academic researchers or tool developers) to understand or take advantage of their potential utility. This second direction aims to make improvements in how we document our outputs (e.g. detailed data statements for corpora we release, targeted literature reviews for specific research areas), and in the development of schemas and data structures associated with that output (e.g. better structured content data dumps, richer and comprehensive cross-language alignments, etc.).

Proposed objectives for this research direction are the following

2.1 Develop new models to help researchers and developers interact with Wikimedia content

- Develop new models to represent knowledge in Wikipedia articles and categories, as well as Wikidata items or entries in other sister projects, in a machine-readable and human-interpretable format. For example, topic models for Wikipedia articles.

2.2 Manage Wikipedia content across languages as a single entity, rather than language specific entities

- We'd like to be able to link a Wikipedia article or concept from one language to an article, section of an article, or a set of articles that describe the same concept in any other language. Thus, machine representations of content (e.g., topic models, embeddings) can be language-independent.

2.3 Produce richer data structures of Wikimedia contents

- Structured data dumps of Wikimedia contents, data and metadata can massively improve how efficiently they can be used by researchers, tool developers and other parties. Rich data dumps may include, for example, article contents in plain text format and section titles, or datasets for dedicated resources such as links, images, and citations.
- Dumps are not the only data structure that can better support research and development on Wikimedia projects: the same considerations apply to new types of real-time data streams (such as those built on top of the EventStream platform) allowing the creation of incremental dumps and the subscription to specific types of events.
- Processing Wikimedia dumps is currently possible only by parties that dispose of significant computational resources. The availability of new data structures would enable researchers to focus on research questions, rather than pre-processing data into the right format.
- These data structures would also support reproducible research as the canonical source of information across the research community which, in turn, can improve the accessibility and reuse of Wikimedia data beyond our immediate ecosystem.

3. Develop principles for ethical AI and responsible research practices for WMF and the Wikimedia Movement

As a technology organization at the service of a community of volunteers and as an organization with a free culture ethos, a mandate to publish openly, and a commitment to pursuing a research agenda for social good, we have an opportunity to serve as a leader in what it means to do human-centered data science and ethical AI. We should articulate and promote our principles. We should develop and adhere to processes that reflects those principles. We should hold ourselves publicly accountable for outcomes of our work.

The proposed objectives for this third research direction are the following:

3.1 Develop and promote an ethical AI evaluation strategy

- Develop standards and pilot new methods for evaluating AI technologies. This might include strategies such as checklists and benchmark datasets against which to evaluate our AI technologies. We should ensure that these approaches are effective for promoting fairness and provide useful transparency around our technologies.
- Share and promote standards for fair and transparent AI with the broader scientific community and technology industry, via academic publications and presence in related outreach events.

3.2 Develop AI technologies that are inclusive by design

- Our technologies are meant to support multilingual and culturally diverse communities. We should take into account this diversity throughout the process of creation of our AI technologies, from concept design to productization and evaluation, avoiding or improving solutions that could be tailored to dominant culture/languages only.

3.3 Develop and promote open AI technologies

- As an organization, we believe in free, open access to knowledge. The technologies we develop should reflect this principle. We shall develop free, open AI technologies that are based on open source tools and components, and make our code, data and results publicly available.

3.4 Inspired by Communities

- As researchers at the Wikimedia Foundation, we are inspired by the broad community of volunteers in Wikimedia spaces, and by the latest technology produced by our scientific communities. Part of our job is to bridge these two communities, and prioritize those types of AI technologies that can help Wikimedians improve global access to free knowledge.

4. Build research capacity and outreach in the movement

The Wikimedia Movement is a global movement and no one team or organization can represent and satisfy all the needs of the movement as a whole. We have an opportunity to rethink the way in which we do research across the globe and how we support or initiate research communities at a global scale.

There are multiple opportunities and avenues we can explore, including building a stronger and more diverse collaboration network on top of our existing Formal Collaboration program; build partnerships with research institutions around the world, especially in regions where we do not have strong research presence today; expand on the current Research Fellows program to work more closely with researchers in the Wikimedia field; develop and grow an internship program; work more closely with grant making teams for mentoring research projects, and working with Wikimedia communities of volunteer editors, organizers, and affiliates to help bridge the gap between research needs and research resources in the corresponding communities. We intend to help build research capacity in the movement through the following opportunities:

4.1 Enhance the discoverability and applicability of research results by Wikimedia communities

- Design a central landing page for Wikimedia research, including live documentation summarizing the literature and state of the art on different topics, data sources available, frequently asked questions for researchers working with Wikimedia data, etc.
- Develop community feedback or wishlist mechanisms for research, allowing groups and communities to submit and prioritize questions researchers at the Wikimedia Foundation and in the academia community can focus on.

4.2 Provide structured outreach and build community capacity

- Develop replicable modules for research outreach.
- Engage students and researchers that are not experts in Wikimedia projects, but have a desire to get involved via mentorship, internship programs, etc.
- Create a stronger Research Fellow program by expanding the current

program to include more researchers from across the globe with different sets of expertise. Identify and promote the young researchers across the globe who can become champions of Wikimedia research, using models that have already been designed in other areas of the movement.

- Expand the Wikimedia Foundation's Formal Collaboration program from tens of researchers to hundreds of researchers across the globe and repurpose it to bridge the gaps between the community of Wikimedia volunteer contributors and developers and academic researchers.