Building open tools to support research on Wikimedia projects

Martin Gerlach, Emily Lescak, Pablo Aragón
The Wikimedia Foundation Research Team
FOSDEM 2023
It is a non-profit organization of ~600 staff

It provides broad support to Wikimedia communities and projects: servers, data centers, legal and communications support, etc.

It does not create or modify content.

It does not define or enforce policies on the projects.
Research Team

Pablo Aragón  Martin Gerlach  Isaac Johnson  Fabian Kaelin

Emily Lescak  Miriam Redi  Diego Sáez-Trumper  Leila Zia

3 contractors, 1 Research Fellow

17 Formal Collaborators
Our Programs

Address Knowledge Gaps

Improve Knowledge Integrity

Grow the Research Community
A sustainable distributed network of Wikimedia projects relies on an empowered global community of Wikimedia researchers...
Open Datasets and Challenges
Wikimedia has a lot of data...

About Wikimedia Dumps

Wikimedia provides public dumps of our wikis' content and of related data such as search indexes and short url mappings. The dumps are used by researchers and in offline reader projects, for archiving, for bot editing of the wikis, and for provision of the data in an easily queryable format, among other things. The dumps are free to download and reuse.

https://meta.wikimedia.org/wiki/Data_dumps
Wikipedia Image Caption competition

Kaggle-competition: 105 participants, open source solutions: https://www.kaggle.com/c/wikipedia-image-caption-caption/overview
And even more open data...

**Predicted quality scores** for all Wikipedia articles:

https://analytics.wikimedia.org/published/datasets/one-off/isaacj/quality/V2_2022_01/README.md

Wikipedia **article readability** data (10+ languages):

https://w.wiki/64CC

Upcoming **differential privacy releases** for reader geography and others:

https://w.wiki/64CE
Data Analysis and Processing Tools
Making data available accessible

New Dump dataset (Oct 2021)

All articles of text-based wikimedia projects (wikipedia, wikisource, etc)

Wikimedia Enterprise HTML Dumps

This partial mirror of Wikimedia Enterprise HTML dumps is an experimental service.

https://dumps.wikimedia.org/other/enterprise_html/
Why are HTML-dumps exciting?

- Traditional dumps contain only the “wikitext” markup of an article
- wikitext gets parsed into HTML (i.e. what a reader sees)
- Problem: wikitext < HTML

Mitrevski et al.: WikiHist.html: English Wikipedia's Full Revision History in HTML Format
Parsing HTML

Challenge: How to (easily) parse the HTML of an article?
Python library for parsing HTML Dumps

Work in progress. Contributions are welcome.

https://gitlab.wikimedia.org/repos/research/html-dumps
Machine Learning APIs
Knowledge integrity

Aim: Support editors in reviewing edits

There are a lot of edits:
~100k edits/day only in English Wikipedia

Asthana&Halfaker: With few eyes, All hoaxes are deep
Revert Risk Model

Binary Classification model: Should the edit be reverted?

Features
Text changed?
Links removed?
Images removed?
Templates removed?
...
Revert Risk Model

Binary Classification model: Should the edit be reverted?

Model performance:
True value: **IS_REVERT**
Predicted value: **IS_REVERT**
**IS_REVERT** predicted probability: 0.992
Some of our ML APIs ...

Quality:
Featured Article

Images:

Geography:
Egypt

ML-Tools for Knowledge Integrity:
"Building a new generation of ML models to support patrolling and anti-vandalism tasks"

Related Articles:
Synesius (disciple)
Theon of Alexandria (father)
Cyril of Alexandria

Topics:
Biography
Philosophy/Religion
STEM

Readability:
Medium
Our Machine Learning models

Open
Training and inference code are open and public

Reliable and Scalable
In collaboration with the ML Platform Team

Multilingual
Preferring Language-Agnostic approaches, to give the same opportunities to all our communities

Explainable
Explainability is as important as accuracy

Community-centered
Communities are encouraged to provide feedback or report biases, to continuously improve models
How to contribute
Discover and build Wikimedia technology

Find technical documentation, and connect with the developer community behind Wikipedia and other Wikimedia projects.

Get started

https://developer.wikimedia.org/

---

Learn how contributing works

Get the basics of how to contribute to MediaWiki and other Wikimedia open source projects.

Read the code of conduct

The code of conduct for Wikimedia technical spaces applies to both physical and virtual environments. Learn how to foster an open and welcoming community.

- Read more on mediawiki.org

Create a developer account

Most Wikimedia technical spaces use a developer account to identify and authorize users. Create a free account to get started as a technical contributor.

- Read more on mediawiki.org

Find projects and tasks for new contributors

Choose from open source projects that offer mentoring for new contributors, and find good first tasks.

- Read more on mediawiki.org

Review guidelines and code conventions

Understand development policies, best practices, and code conventions for Wikimedia software.

- Read more on mediawiki.org
Beyond Wikipedia: Discovering Wikimedia's Open-Source Ecosystem

While the Wikimedia Foundation is best known for its flagship project, Wikipedia, and the MediaWiki software that powers it, the Foundation's open-source ecosystem extends far beyond these well-known projects.

In this talk, we will explore the fascinating world of Wikimedia's open-source tools ecosystem and the cloud infrastructure that makes it possible. We will showcase some of the coolest tools and projects, and we will highlight the unique opportunity that the Foundation offers for contributing to its cloud infrastructure - a rare chance to work on infrastructure for a cause that does good in the world, supporting the Foundation's mission of providing free and open knowledge to the global community.

Whether you are a seasoned open-source developer or a newcomer to the field, this talk will provide valuable insights and inspiration for getting involved in Wikimedia's vibrant community of contributors.

Speakers
Slavina Stefanova

[https://fosdem.org/2023/schedule/event/beyond_wikipedia/](https://fosdem.org/2023/schedule/event/beyond_wikipedia/)
Do you want to ...

build a tool to help Wikipedians?

- Host it on ToolForge [https://wikitech.wikimedia.org/wiki/Portal:Toolforge](https://wikitech.wikimedia.org/wiki/Portal:Toolforge)

---

**Toolforge** is a hosting environment, also known as Platform as a Service. Toolforge makes it easy for you to perform analytics, administer bots, run webservices, and create tools. Tools help project editors, technical contributors, and other volunteers who work on Wikimedia projects.

Toolforge is part of the Wikimedia Cloud Services (WMCS) suite of services. It is supported by Wikimedia Foundation staff and volunteers.

Create or deploy your own tools on Toolforge [edit | edit source]

Toolforge quickstart Create and manage tool accounts
Do you want to ... 

reuse / fix / improve our tools and algorithms?

- Check our repositories/packages

```
[mwparserfromhtml 0.0.5]
pip install mwparserfromhtml
Released: Sep 27, 2022
```

```
[mwedittypes 2.0.2]
pip install mwedittypes
Released: Dec 7, 2022
```

```
[mwsq| 0.1.5]
pip install mwsq
Released: Jan 31, 2022
```

https://gitlab.wikimedia.org/repos/research
Do you want to ... get funding?

- Apply for the Wikimedia Research and Technology Fund
  [https://w.wiki/4LSP](https://w.wiki/4LSP)
Do you want to …

learn more about our research projects?

- Read our research report [https://research.wikimedia.org/report.html](https://research.wikimedia.org/report.html)

---

Research Report Nº 7

December 14, 2022

The seventh in a series of biannual reports from Wikimedia Research, published every June and December.

Executive Summary

Welcome! We are the Wikimedia Foundation’s Research team. We turn research questions into publicly shared knowledge. We design and test new technologies, produce empirical insights to support new products and programs, and publish research that informs the Wikimedia Foundation’s and the Movement’s strategy. We help to build a strong and diverse community of Wikimedia researchers globally. This Research Report is an overview of our team’s latest developments — an entry point that highlights existing and new work, and details new collaborations and considerations, including trends that we’re watching.
Do you want to ...

engage with the Wikimedia Research community?

- Join us in WikiWorkshop [https://wikiworkshop.org](https://wikiworkshop.org)
Thank you

Stay in touch

Reach out mgerlach@wikimedia.org

Office hours https://w.wiki/uJo

Mailing list research-wmf@lists.wikimedia.org

@WikiResearch
@wikiresearch@mastodon.social

IRC #wikimedia-research (libera)

Monthly showcases https://w.wiki/uJn