Building open tools to support research on Wikimedia projects

Martin Gerlach, Emily Lescak, Pablo Aragón
The Wikimedia Foundation Research Team

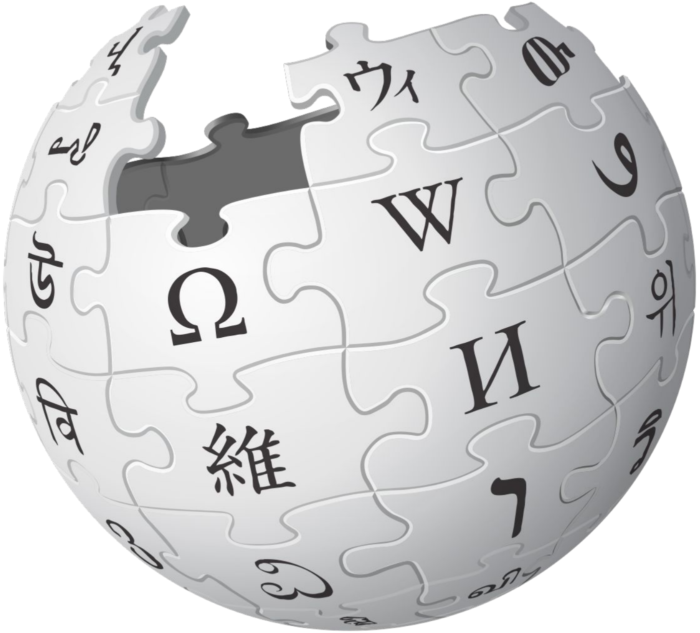
FOSDEM 2023

WIKIMEDIA
FOUNDATION

The largest encyclopedia

56M articles

280+ languages



10M monthly edits

15B monthly pageviews

0.5M volunteer editors

Wikimedia Foundation

- 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

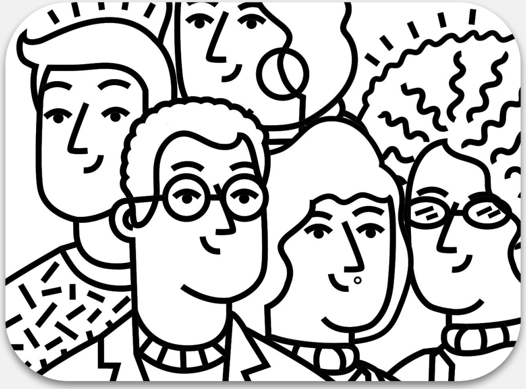
Fabian Kaelin

Emily Lescak

Miriam Redi

Diego Sáez-Trumper

Leila Zia



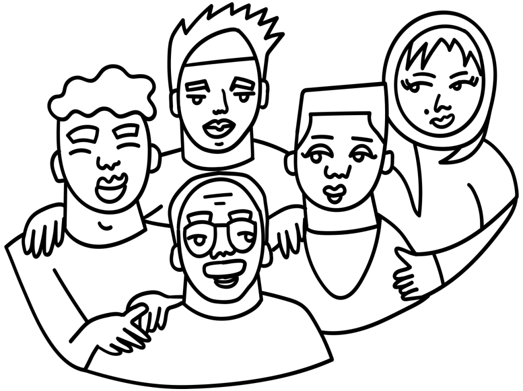
3 contractors, 1 Research Fellow

17 Formal Collaborators

Our Programs

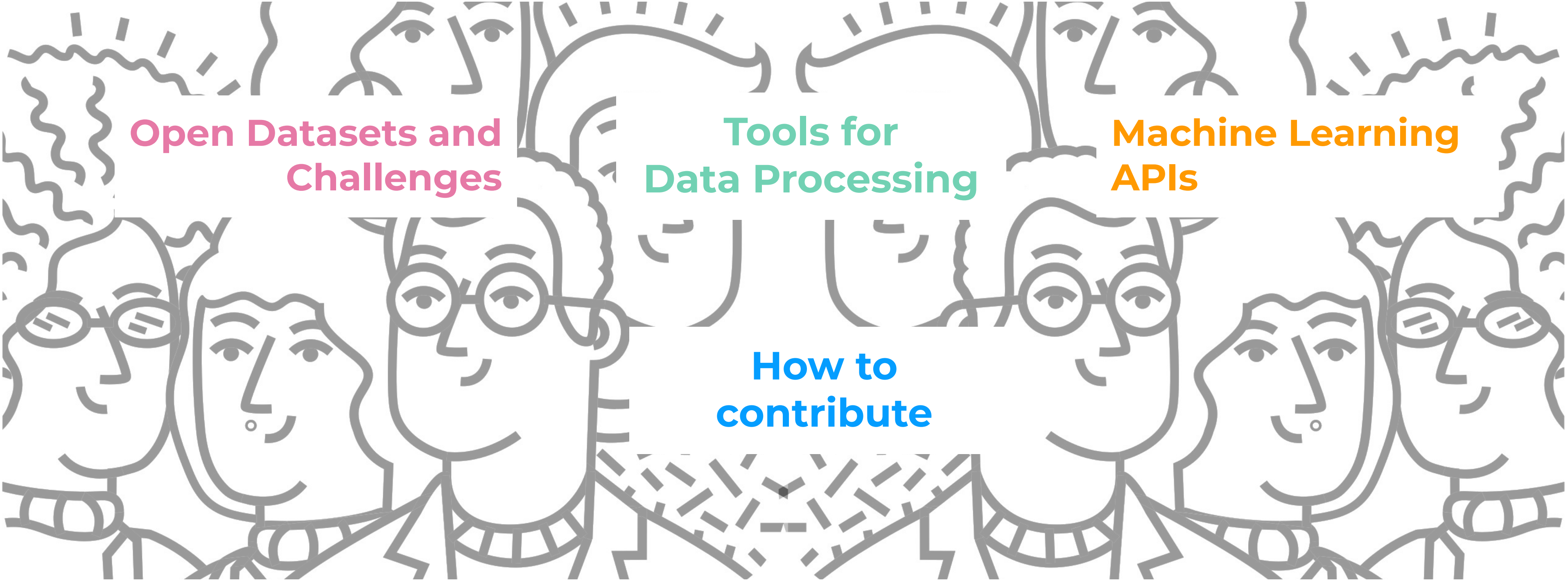
Address Knowledge Gaps

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

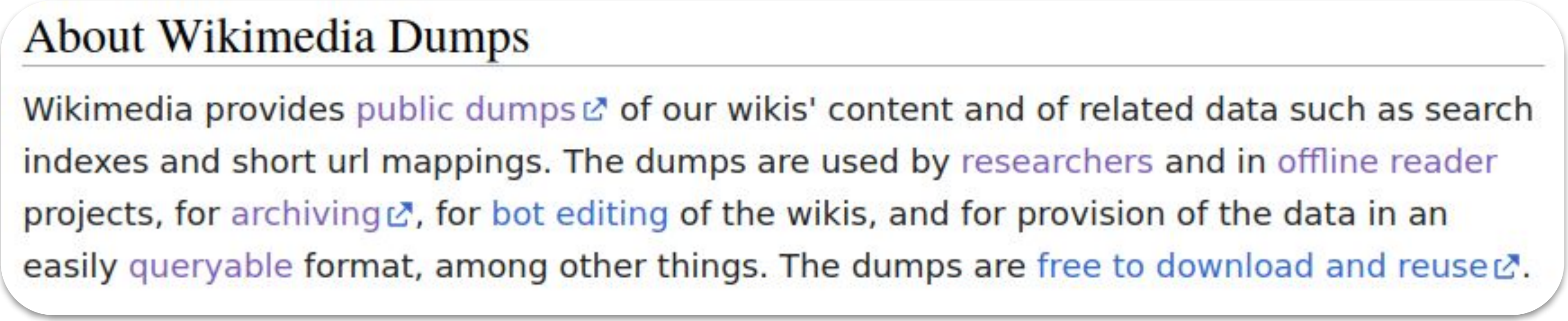
Tools for Data Processing

Machine Learning APIs

How to contribute

Open Datasets and Challenges

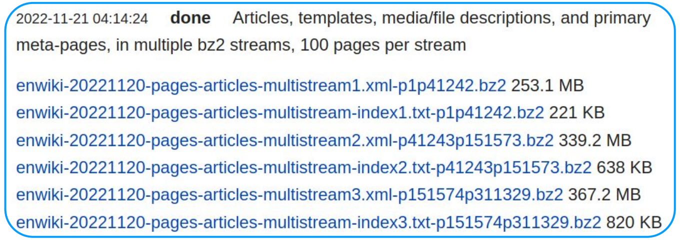
Wikimedia has a lot of data...



About Wikimedia Dumps

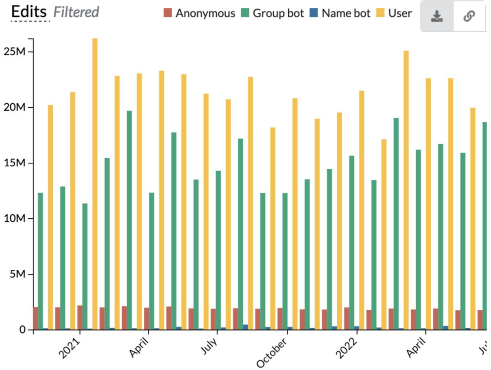
Wikimedia provides public dumps of our wikis' content and of related data such as search indexes and short url mappings. The dumps are used by researchers and in offline reader projects, for archiving, for bot editing of the wikis, and for provision of the data in an easily queryable format, among other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

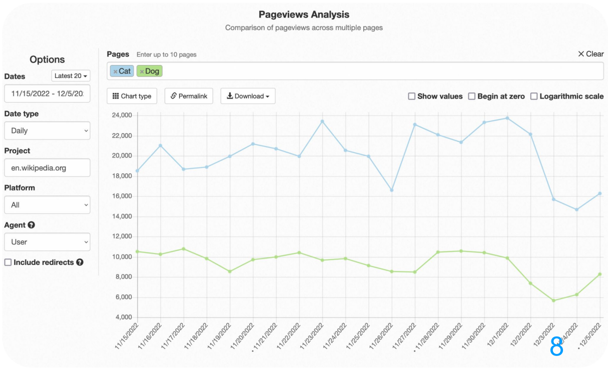
https://meta.wikimedia.org/wiki/Data_dumps



2022-11-21 04:14:24 **done** Articles, templates, media/file descriptions, and primary meta-pages, in multiple bz2 streams, 100 pages per stream

enwiki-20221120-pages-articles-multistream1.xml-p1p41242.bz2 253.1 MB
enwiki-20221120-pages-articles-multistream-index1.txt-p1p41242.bz2 221 KB
enwiki-20221120-pages-articles-multistream2.xml-p41243p151573.bz2 339.2 MB
enwiki-20221120-pages-articles-multistream-index2.txt-p41243p151573.bz2 638 KB
enwiki-20221120-pages-articles-multistream3.xml-p151574p311329.bz2 367.2 MB
enwiki-20221120-pages-articles-multistream-index3.txt-p151574p311329.bz2 820 KB





Pageviews Analysis

Comparison of pageviews across multiple pages

Options

Dates: Latest 20 ▾ — 11/15/2022 - 12/5/20

Date type: Daily

Project: en.wikipedia.org

Platform: All

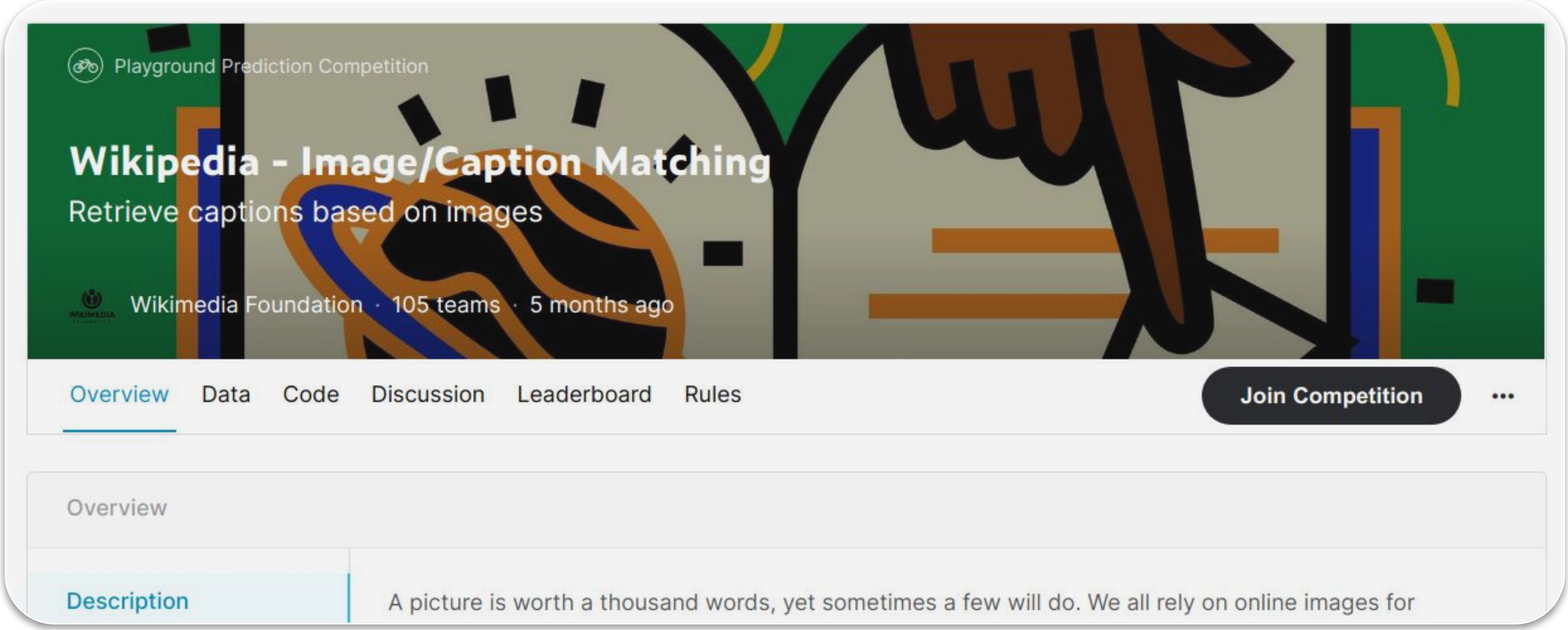
Agent: User

☐ Include redirects

Pages: Enter up to 10 pages — × Cat × Dog — ✕ Clear

☐ Show values ☐ Begin at zero ☐ Logarithmic scale

Wikipedia Image Caption competition



Playground Prediction Competition

Wikipedia - Image/Caption Matching

Retrieve captions based on images

Wikimedia Foundation · 105 teams · 5 months ago

A picture is worth a thousand words, yet sometimes a few will do. We all rely on online images for

Kaggle-competition: 105 participants, open source solutions: https://www.kaggle.com/c/wikipedia-image-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:

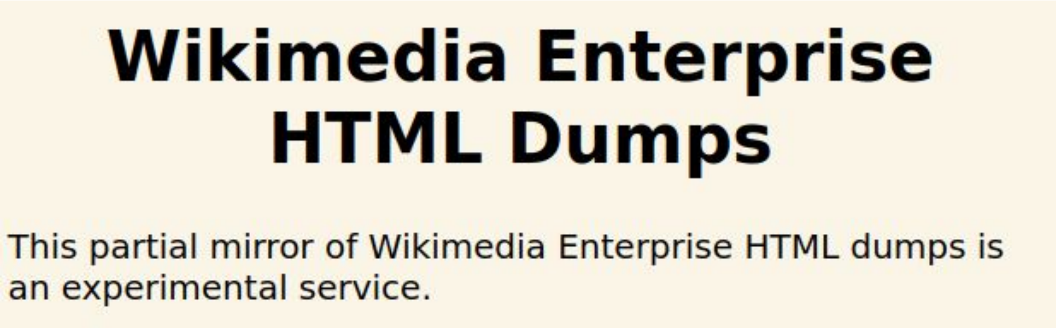
https://w.wiki/64CE

Data Analysis and Processing Tools

Making data ~~available~~ accessible

New Dump dataset (Oct 2021)

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



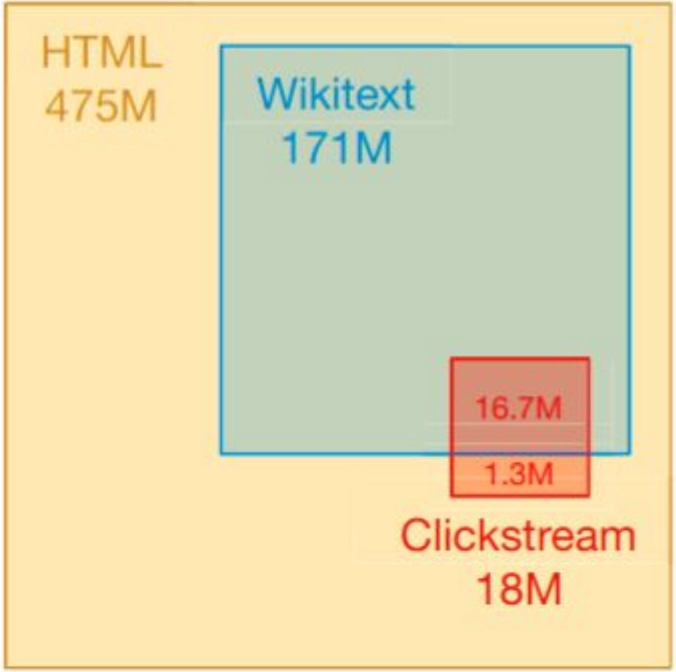
Wikimedia Enterprise HTML Dumps

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

https://dumps.wikimedia.org/other/enterprise_html/

Why are HTML-dumps exciting?

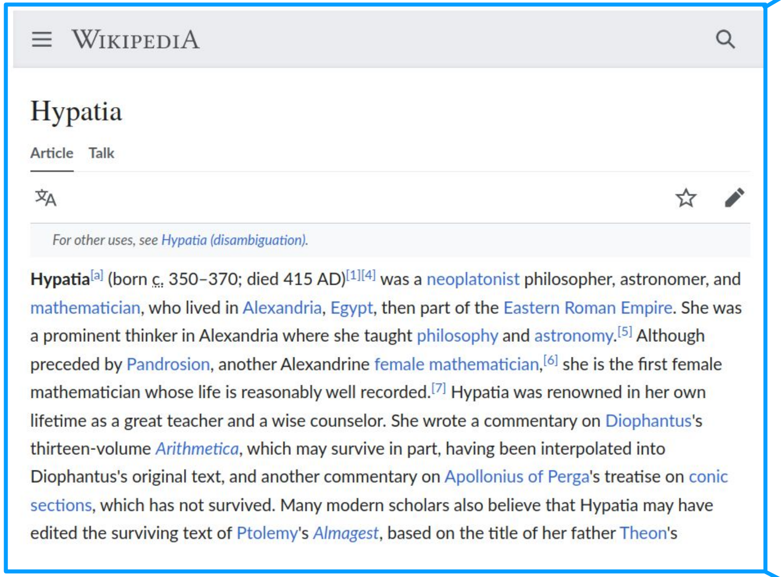
- Traditional dumps contain only the “wikitext” markup of an article
- wikitext gets parsed into HTML (i.e. what a reader sees)
- Problem: wikitext < HTML



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

Parsing HTML

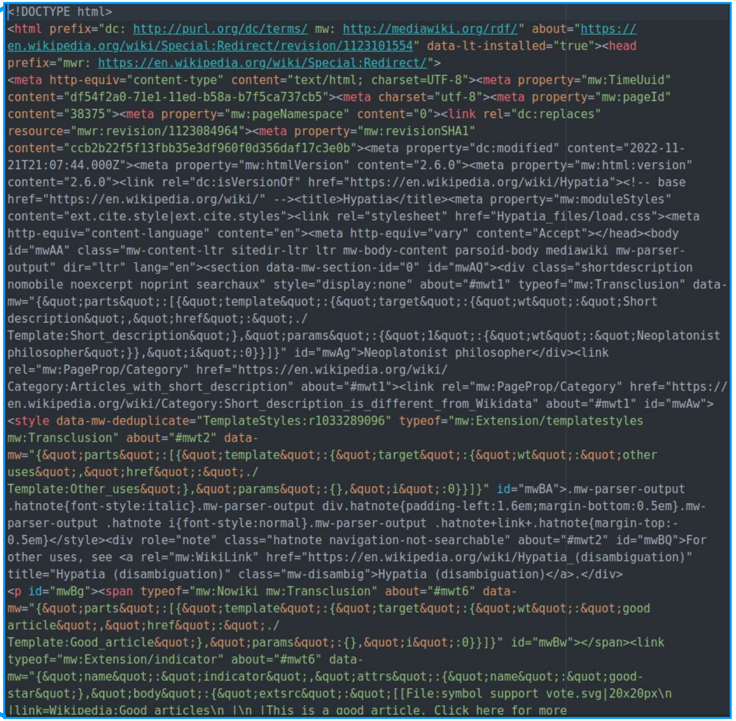
Challenge: How to (easily) parse the HTML of an article?



Hypatia

For other uses, see Hypatia (disambiguation).

Hypatia[a] (born c. 350–370; died 415 AD)[1][4] was a neoplatonist philosopher, astronomer, and mathematician, who lived in Alexandria, Egypt, then part of the Eastern Roman Empire. She was a prominent thinker in Alexandria where she taught philosophy and astronomy.[5] Although preceded by Pandrosion, another Alexandrine female mathematician,[6] she is the first female mathematician whose life is reasonably well recorded.[7] Hypatia was renowned in her own lifetime as a great teacher and a wise counselor. She wrote a commentary on Diophantus's thirteen-volume *Arithmetica*, which may survive in part, having been interpolated into Diophantus's original text, and another commentary on Apollonius of Perga's treatise on conic sections, which has not survived. Many modern scholars also believe that Hypatia may have edited the surviving text of Ptolemy's *Almagest*, based on the title of her father Theon's



```
<!DOCTYPE html>
<html prefix="dc: http://purl.org/dc/terms/ mw: http://mediawiki.org/rdf/" about="https://
en.wikipedia.org/wiki/Special:Redirect/revision/1123101554" data-lt-installed="true"><head
prefix="mwr: https://en.wikipedia.org/wiki/Special:Redirect/">
<meta http-equiv="content-type" content="text/html; charset=UTF-8"><meta property="mw:TimeUuid"
content="df54f2a0-71e1-11ed-b58a-b7f5ca737cb5"><meta charset="utf-8"><meta property="mw:pageId"
content="38375"><meta property="mw:pageNamespace" content="0"><link rel="dc:replaces"
resource="mwr:revision/1123084964"><meta property="mw:revisionSHA1"
content="ccb2b22f5f13fbb35e3df960f0d356daf17c3e0b"><meta property="dc:modified" content="2022-11-
21T21:07:44.000Z"><meta property="mw:htmlVersion" content="2.6.0"><meta property="mw:html:version"
content="2.6.0"><link rel="dc:isVersionOf" href="https://en.wikipedia.org/wiki/Hypatia"><!-- base
href="https://en.wikipedia.org/wiki/" --><title>Hypatia</title><meta property="mw:moduleStyles"
content="ext.cite.style|ext.cite.styles"><link rel="stylesheet" href="Hypatia_files/load.css"><meta
http-equiv="content-language" content="en"><meta http-equiv="vary" content="Accept"></head><body
id="mwAA" class="mw-content-ltr sitedir-ltr ltr mw-body-content parsoid-body mediawiki mw-parser-
output" dir="ltr" lang="en"><section data-mw-section-id="0" id="mwAQ"><div class="shortdescription
nomobile noexcerpt noprint searchaux" style="display:none" about="#mwt1" typeof="mw:Transclusion" data-
mw="{&quot;parts&quot;:[{&quot;template&quot;:{&quot;target&quot;:{&quot;wt&quot;:&quot;Short
description&quot;,&quot;href&quot;:&quot;./
Template:Short_description&quot;},&quot;params&quot;:{&quot;1&quot;:{&quot;wt&quot;:&quot;Neoplatonist
philosopher&quot;}},&quot;i&quot;:0}}]}" id="mwAg">Neoplatonist philosopher</div><link
rel="mw:PageProp/Category" href="https://en.wikipedia.org/wiki/
Category:Articles_with_short_description" about="#mwt1"><link rel="mw:PageProp/Category" href="https://
en.wikipedia.org/wiki/Category:Short_description_is_different_from_Wikidata" about="#mwt1" id="mwAw">
<style data-mw-deduplicate="TemplateStyles:r1033289096" typeof="mw:Extension/templatestyles
mw:Transclusion" about="#mwt2" data-
mw="{&quot;parts&quot;:[{&quot;template&quot;:{&quot;target&quot;:{&quot;wt&quot;:&quot;other
uses&quot;,&quot;href&quot;:&quot;./
Template:Other_uses&quot;},&quot;params&quot;:{},&quot;i&quot;:0}}]}" id="mwBA">.mw-parser-output
.hatnote{font-style:italic}.mw-parser-output div.hatnote{padding-left:1.6em;margin-bottom:0.5em}.mw-
parser-output .hatnote i{font-style:normal}.mw-parser-output .hatnote+link+.hatnote{margin-top:-
0.5em}</style><div role="note" class="hatnote navigation-not-searchable" about="#mwt2" id="mwBQ">For
other uses, see <a rel="mw:WikiLink" href="https://en.wikipedia.org/wiki/Hypatia_(disambiguation)"
title="Hypatia (disambiguation)" class="mw-disambig">Hypatia (disambiguation)</a>.</div>
<p id="mwBg"><span typeof="mw:Nowiki mw:Transclusion" about="#mwt6" data-
mw="{&quot;parts&quot;:[{&quot;template&quot;:{&quot;target&quot;:{&quot;wt&quot;:&quot;good
article&quot;,&quot;href&quot;:&quot;./
Template:Good_article&quot;},&quot;params&quot;:{},&quot;i&quot;:0}}]}" id="mwBw"></span><link
typeof="mw:Extension/indicator" about="#mwt6" data-
mw="{&quot;name&quot;:&quot;indicator&quot;,&quot;attrs&quot;:{&quot;name&quot;:&quot;good-
star&quot;},&quot;body&quot;:{&quot;extsrc&quot;:&quot;[[File:symbol support vote.svg|20x20px\n
|link=Wikipedia:Good articles\n |\n |This is a good article. Click here for more
```

https://wiki-topic.toolforge.org/html-parsing

mwparserfromhtml 0.0.5

```
pip install mwparserfromhtml
```

Released: Sep 27, 2022

Python library for parsing HTML Dumps

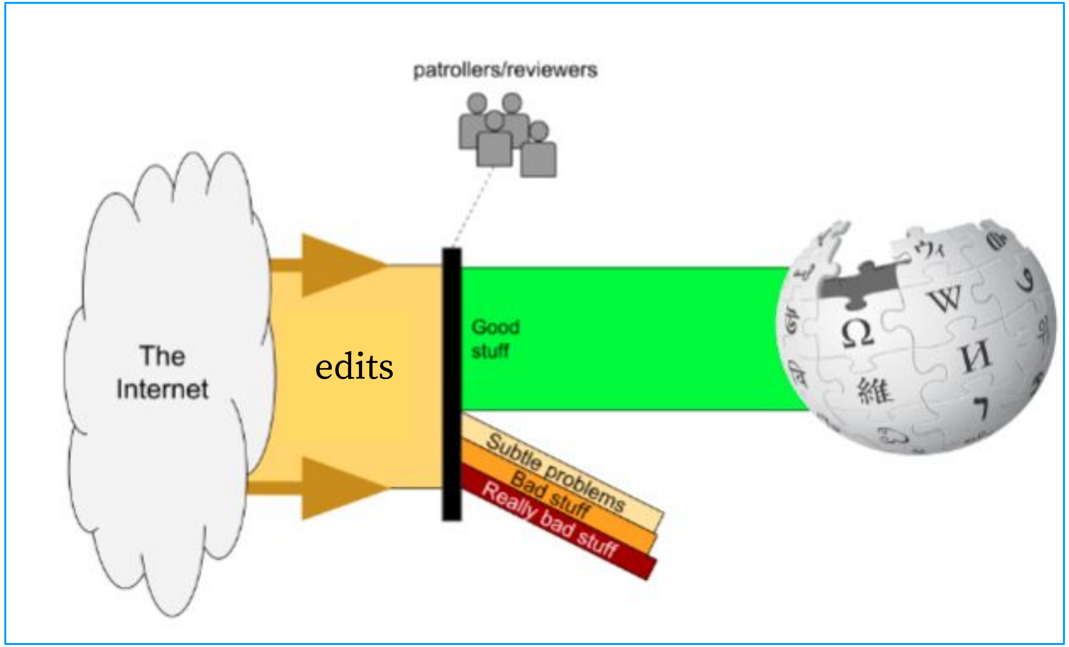
Work in progress. Contributions are welcome.

https://gitlab.wikimedia.org/repos/research/html-dumps

Machine Learning APIs

Knowledge integrity

Aim: Support editors in reviewing edits



Asthana&Halfaker: *With few eyes, All hoaxes are deep*

There are a lot of edits:

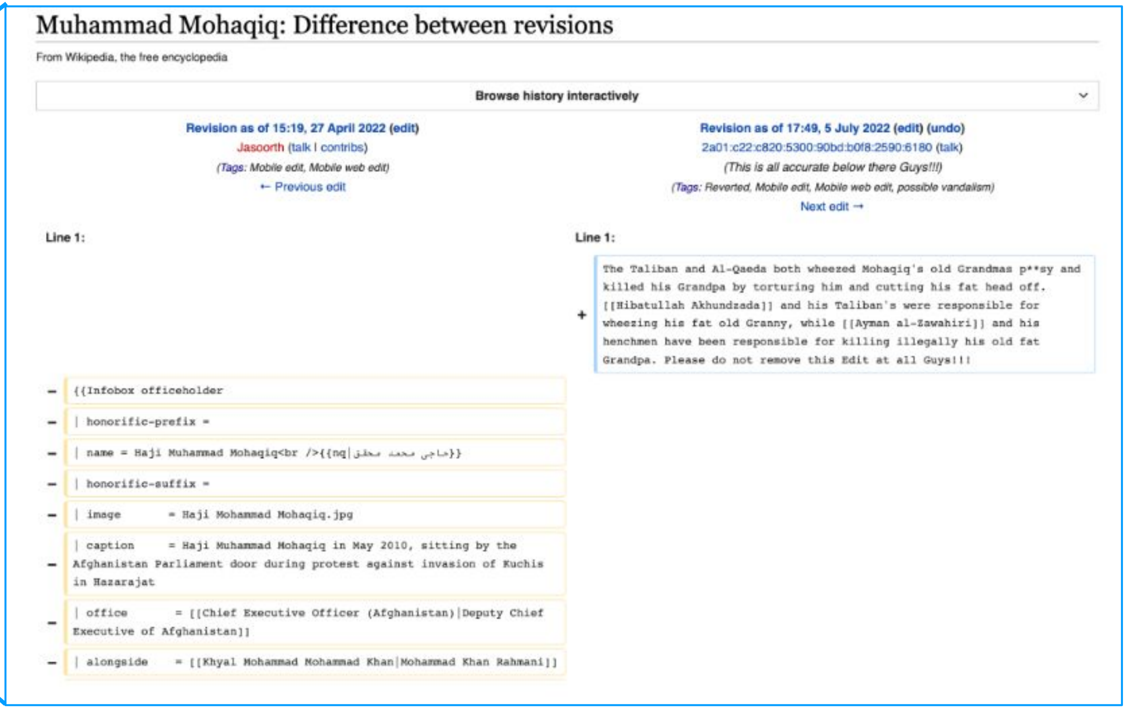
~100k edits/day only in English Wikipedia

Revert Risk Model

Binary Classification model: Should the edit be reverted?

<u>Features</u>

Text changed?
Links removed?
Images removed?
Templates removed?
…



Muhammad Mohaqiq: Difference between revisions

From Wikipedia, the free encyclopedia

Browse history interactively

Revision as of 15:19, 27 April 2022 (edit)
Jasoorth (talk | contribs)
(Tags: Mobile edit, Mobile web edit)
← Previous edit

Revision as of 17:49, 5 July 2022 (edit) (undo)
2a01:c22:c820:5300:90bd:b0f8:2590:6180 (talk)
(This is all accurate below there Guys!!!)
(Tags: Reverted, Mobile edit, Mobile web edit, possible vandalism)
Next edit →

Line 1:

Line 1:

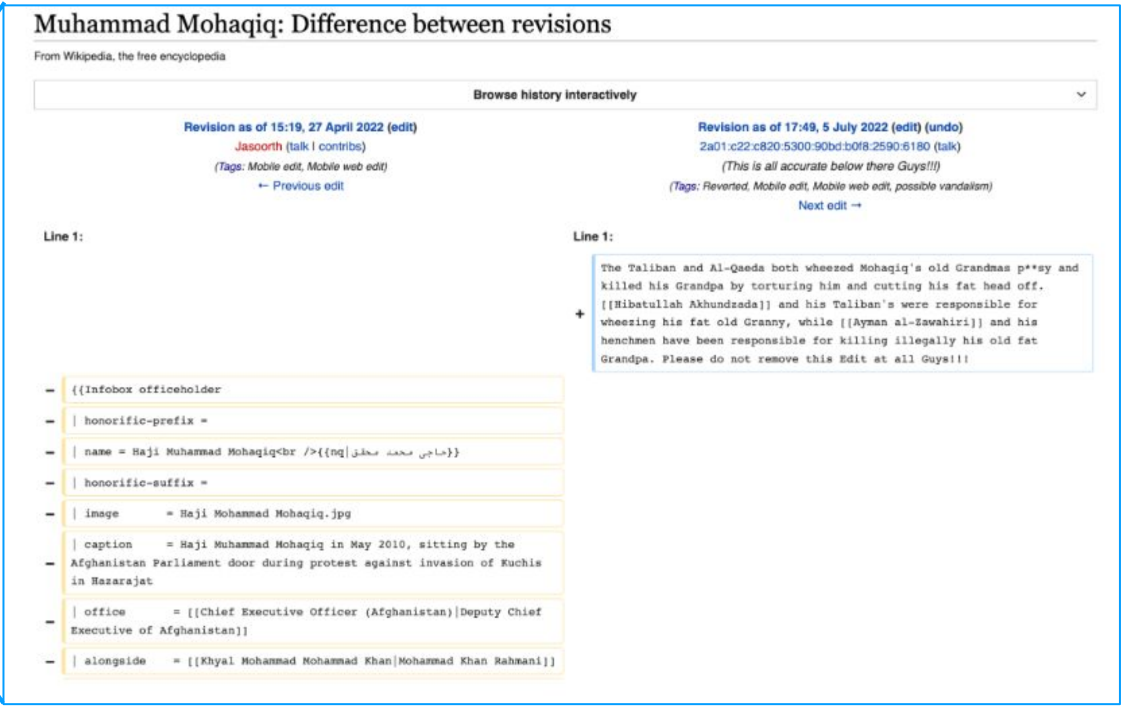
```
+ The Taliban and Al-Qaeda both wheezed Mohaqiq's old Grandmas p**sy and killed his Grandpa by torturing him and cutting his fat head off. [[Hibatullah Akhundzada]] and his Taliban's were responsible for wheezing his fat old Granny, while [[Ayman al-Zawahiri]] and his henchmen have been responsible for killing illegally his old fat Grandpa. Please do not remove this Edit at all Guys!!!
```

```
− {{Infobox officeholder
− | honorific-prefix =
− | name = Haji Muhammad Mohaqiq<br />{{nq|حاجی محمد محقق}}
− | honorific-suffix =
− | image        = Haji Mohammad Mohaqiq.jpg
− | caption      = Haji Muhammad Mohaqiq in May 2010, sitting by the Afghanistan Parliament door during protest against invasion of Kuchis in Hazarajat
− | office        = [[Chief Executive Officer (Afghanistan)|Deputy Chief Executive of Afghanistan]]
− | alongside     = [[Khyal Mohammad Mohammad Khan|Mohammad Khan Rahmani]]
```

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



Muhammad Mohaqiq: Difference between revisions

From Wikipedia, the free encyclopedia

Browse history interactively

Revision as of 15:19, 27 April 2022 (edit)
Jasoorth (talk | contribs)
(Tags: Mobile edit, Mobile web edit)
← Previous edit

Revision as of 17:49, 5 July 2022 (edit) (undo)
2a01:c22:c820:5300:90bd:b0f8:2590:6180 (talk)
(This is all accurate below there Guys!!!)
(Tags: Reverted, Mobile edit, Mobile web edit, possible vandalism)
Next edit →

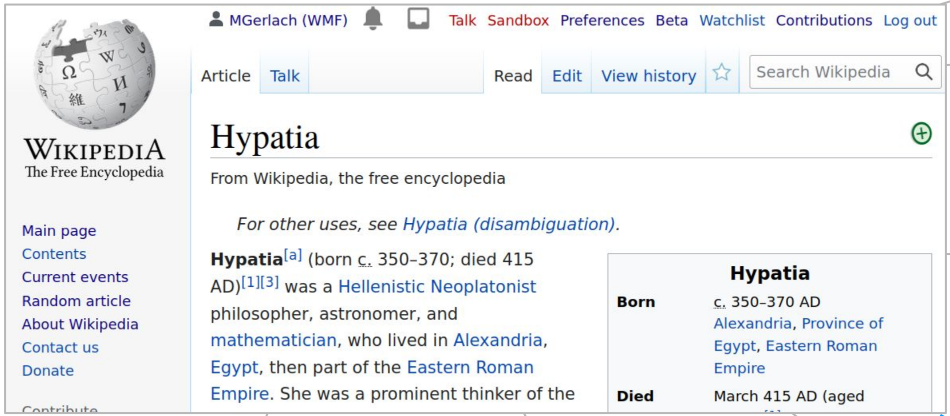
Line 1:

```
− {{Infobox officeholder
− | honorific-prefix =
− | name = Haji Muhammad Mohaqiq<br />{{nq|حاجی محمد محقق}}
− | honorific-suffix =
− | image        = Haji Mohammad Mohaqiq.jpg
− | caption      = Haji Muhammad Mohaqiq in May 2010, sitting by the Afghanistan Parliament door during protest against invasion of Kuchis in Hazarajat
− | office        = [[Chief Executive Officer (Afghanistan)|Deputy Chief Executive of Afghanistan]]
− | alongside     = [[Khyal Mohammad Mohammad Khan|Mohammad Khan Rahmani]]
```

Line 1:

```
+ The Taliban and Al-Qaeda both wheezed Mohaqiq's old Grandmas p**sy and killed his Grandpa by torturing him and cutting his fat head off. [[Hibatullah Akhundzada]] and his Taliban's were responsible for wheezing his fat old Granny, while [[Ayman al-Zawahiri]] and his henchmen have been responsible for killing illegally his old fat Grandpa. Please do not remove this Edit at all Guys!!!
```

Some of our ML APIs ...



MGerlach (WMF) Talk Sandbox Preferences Beta Watchlist Contributions Log out

WIKIPEDIA
The Free Encyclopedia

Article Talk Read Edit View history Search Wikipedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate
Contribute

Hypatia

From Wikipedia, the free encyclopedia

For other uses, see Hypatia (disambiguation).

Hypatia[a] (born c. 350–370; died 415 AD)[1][3] was a Hellenistic Neoplatonist philosopher, astronomer, and mathematician, who lived in Alexandria, Egypt, then part of the Eastern Roman Empire. She was a prominent thinker of the

Hypatia	
Born	c. 350–370 AD Alexandria, Province of Egypt, Eastern Roman Empire
Died	March 415 AD (aged

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

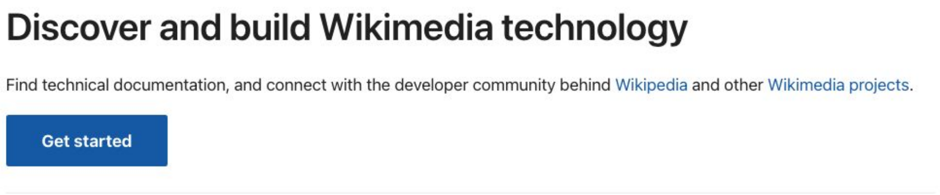
Community-centered

Communities are encouraged to provide feedback or report biases, to continuously improve models

WIKIMEDIA FOUNDATION

How to contribute

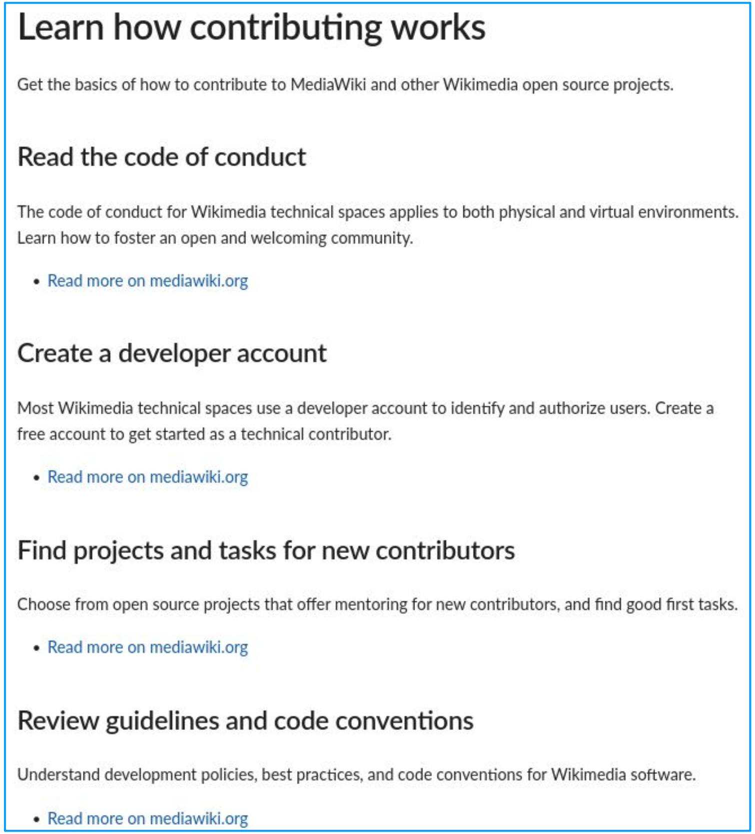
Developer Portal



Discover and build Wikimedia technology

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

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.

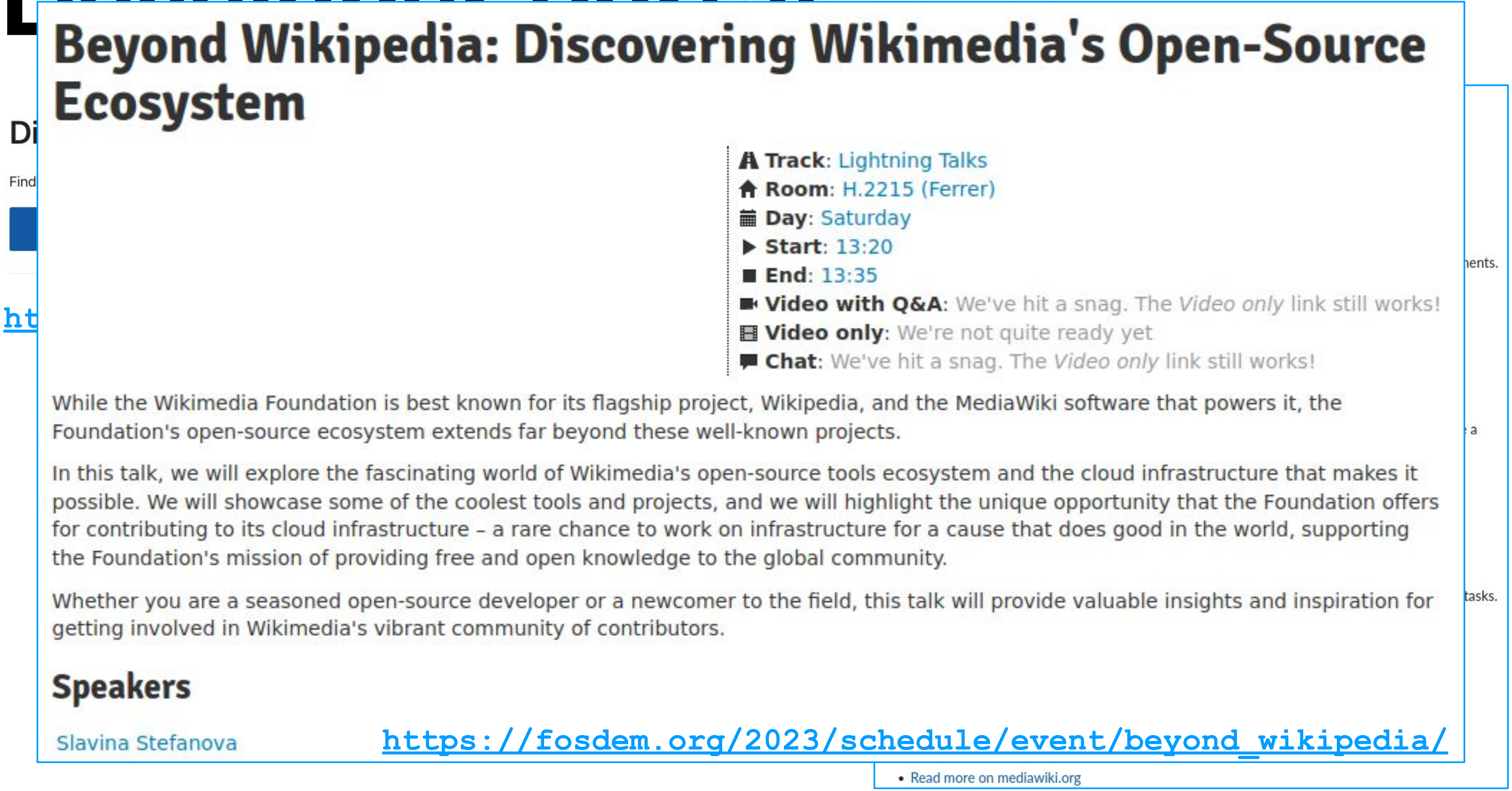
- Read more on mediawiki.org

Review guidelines and code conventions

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

- Read more on mediawiki.org

Developer Portal



Beyond Wikipedia: Discovering Wikimedia's Open-Source Ecosystem

- **Track**: Lightning Talks
- **Room**: H.2215 (Ferrer)
- **Day**: Saturday
- **Start**: 13:20
- **End**: 13:35
- **Video with Q&A**: We've hit a snag. The *Video only* link still works!
- **Video only**: We're not quite ready yet
- **Chat**: We've hit a snag. The *Video only* link still works!

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

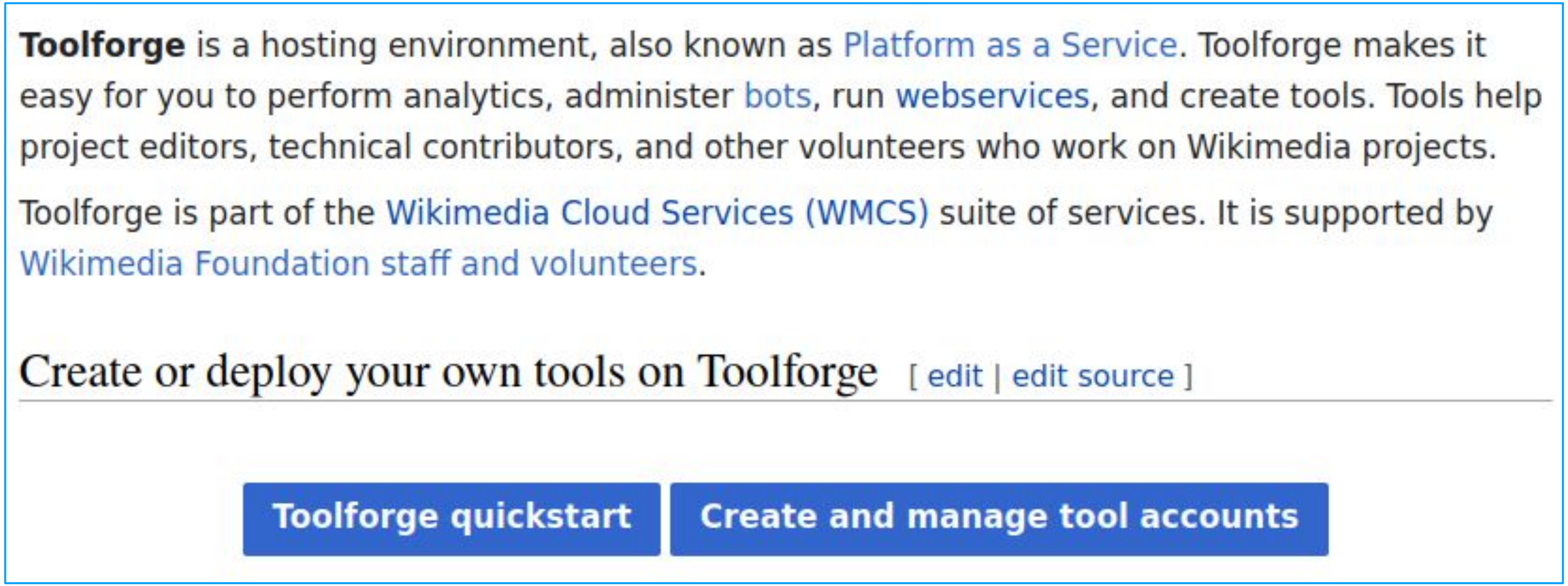
https://fosdem.org/2023/schedule/event/beyond_wikipedia/

- Read more on mediawiki.org

Do you want to ...

build a tool to help Wikipedians?

- Host it on 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.

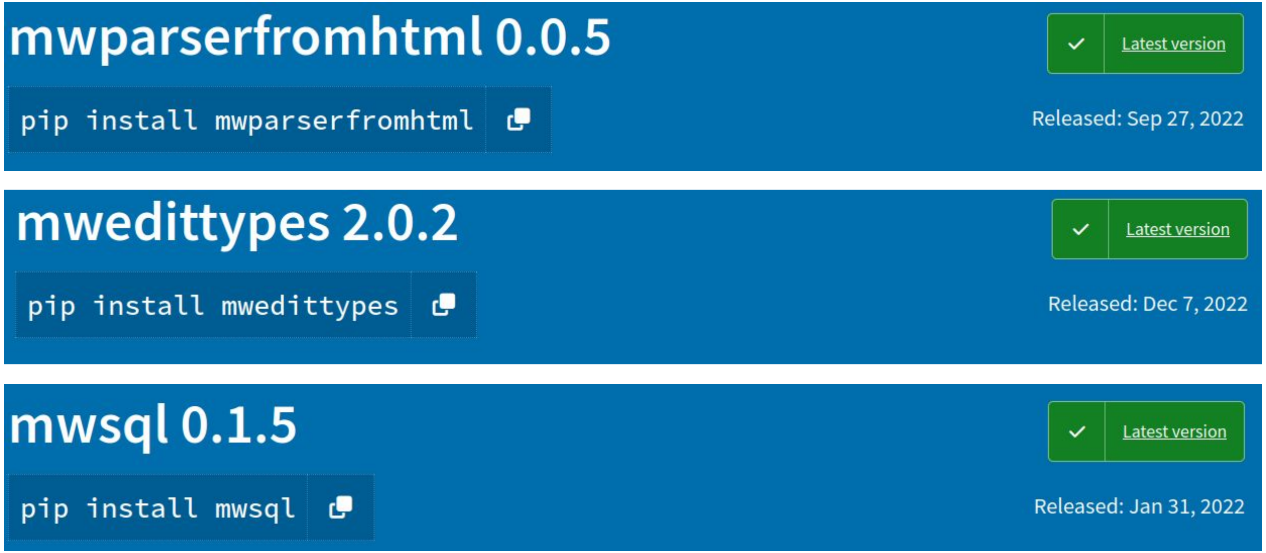
Toolforge is part of the Wikimedia Cloud Services (WMCS) suite of services. It is supported by Wikimedia Foundation staff and volunteers.

Create or deploy your own tools on Toolforge

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

medittypes 2.0.2

```
pip install mwedittypes
```

Released: Dec 7, 2022

mwsql 0.1.5

```
pip install mwsql
```

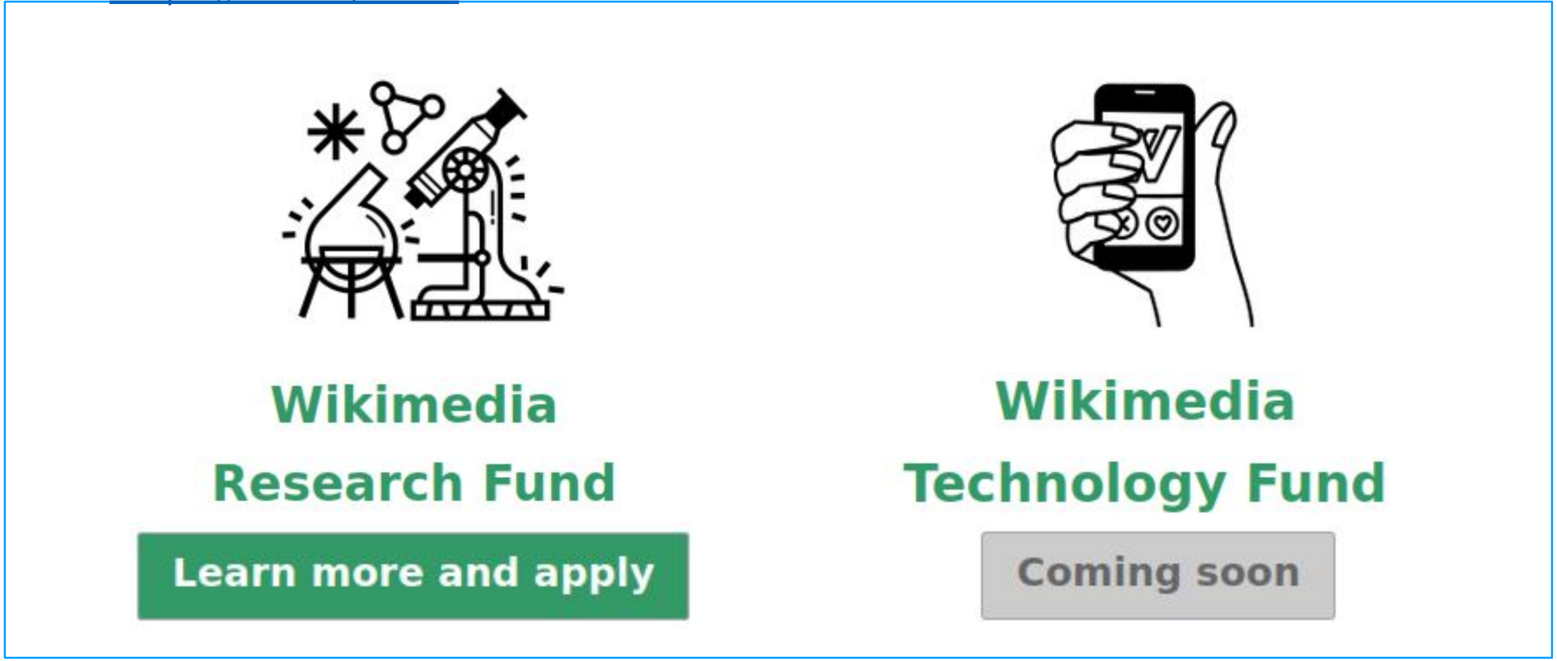
Released: Jan 31, 2022

https://gitlab.wikimedia.org/repos/research

Do you want to …

get funding?

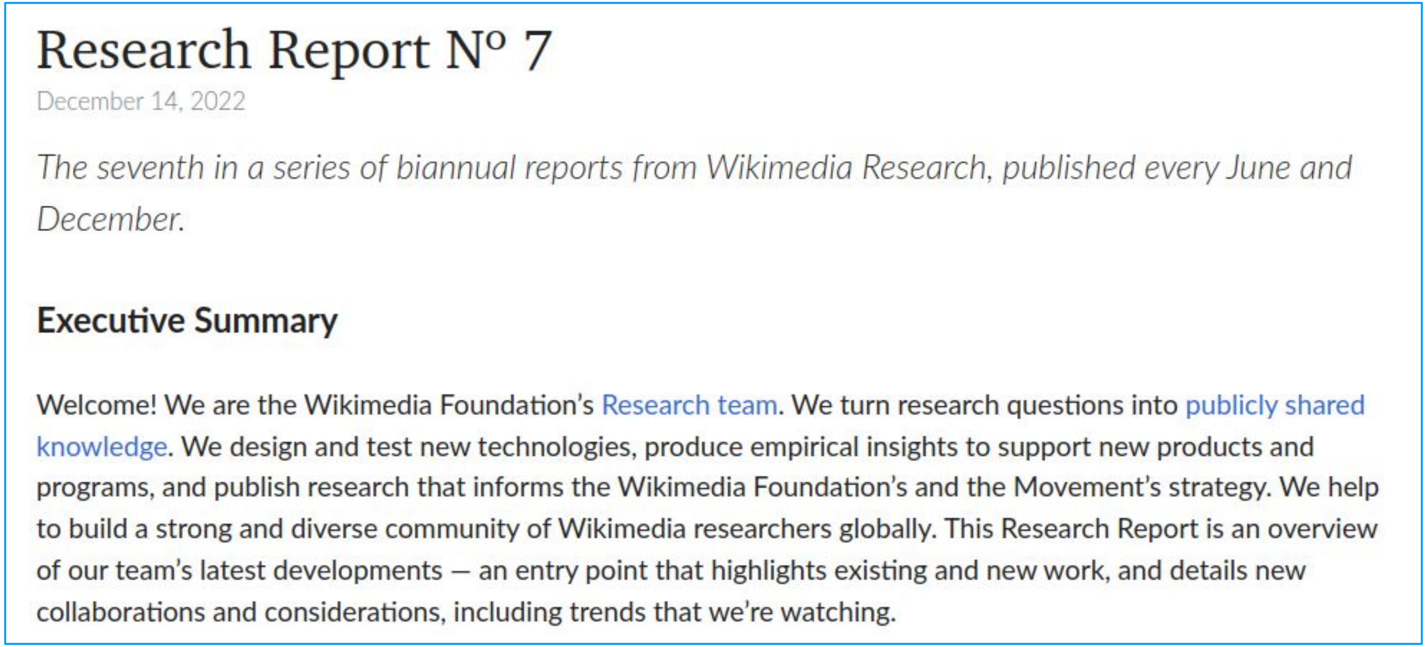
- Apply for the Wikimedia Research and Technology Fund https://w.wiki/4LSP



Do you want to ...

learn more about our research projects?

- Read our research report 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

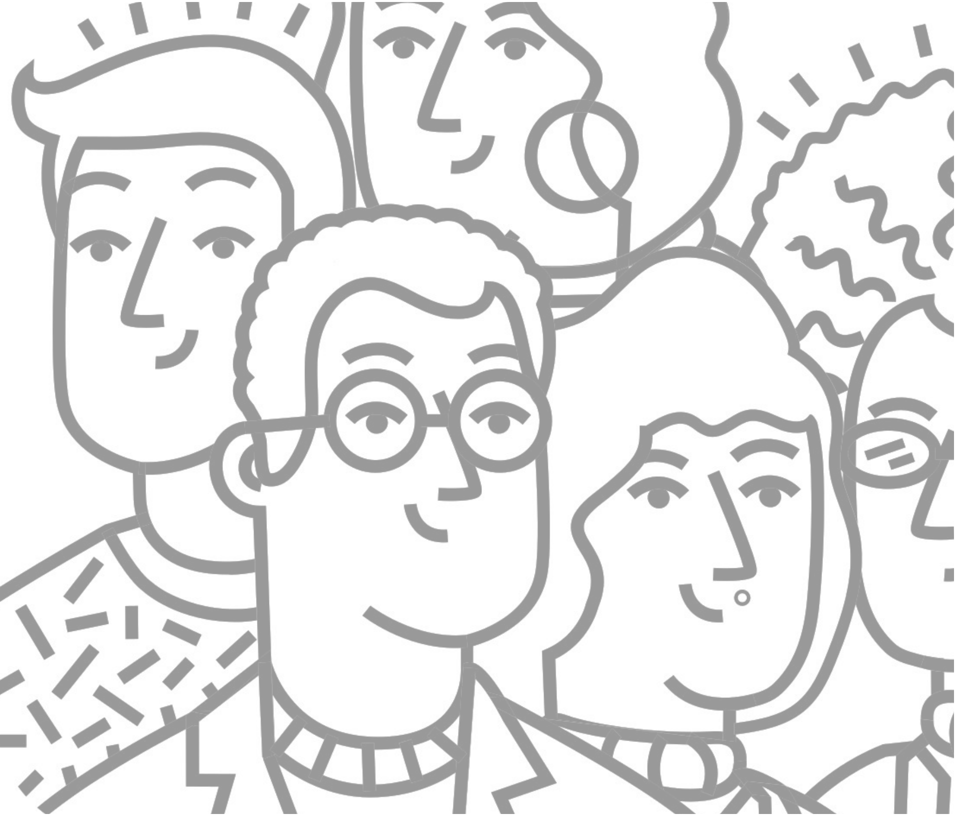
Welcome! We are the Wikimedia Foundation's Research team. We turn research questions into publicly shared knowledge. We design and test new technologies, produce empirical insights to support new products and programs, and publish research that informs the Wikimedia Foundation's and the Movement's strategy. 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

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