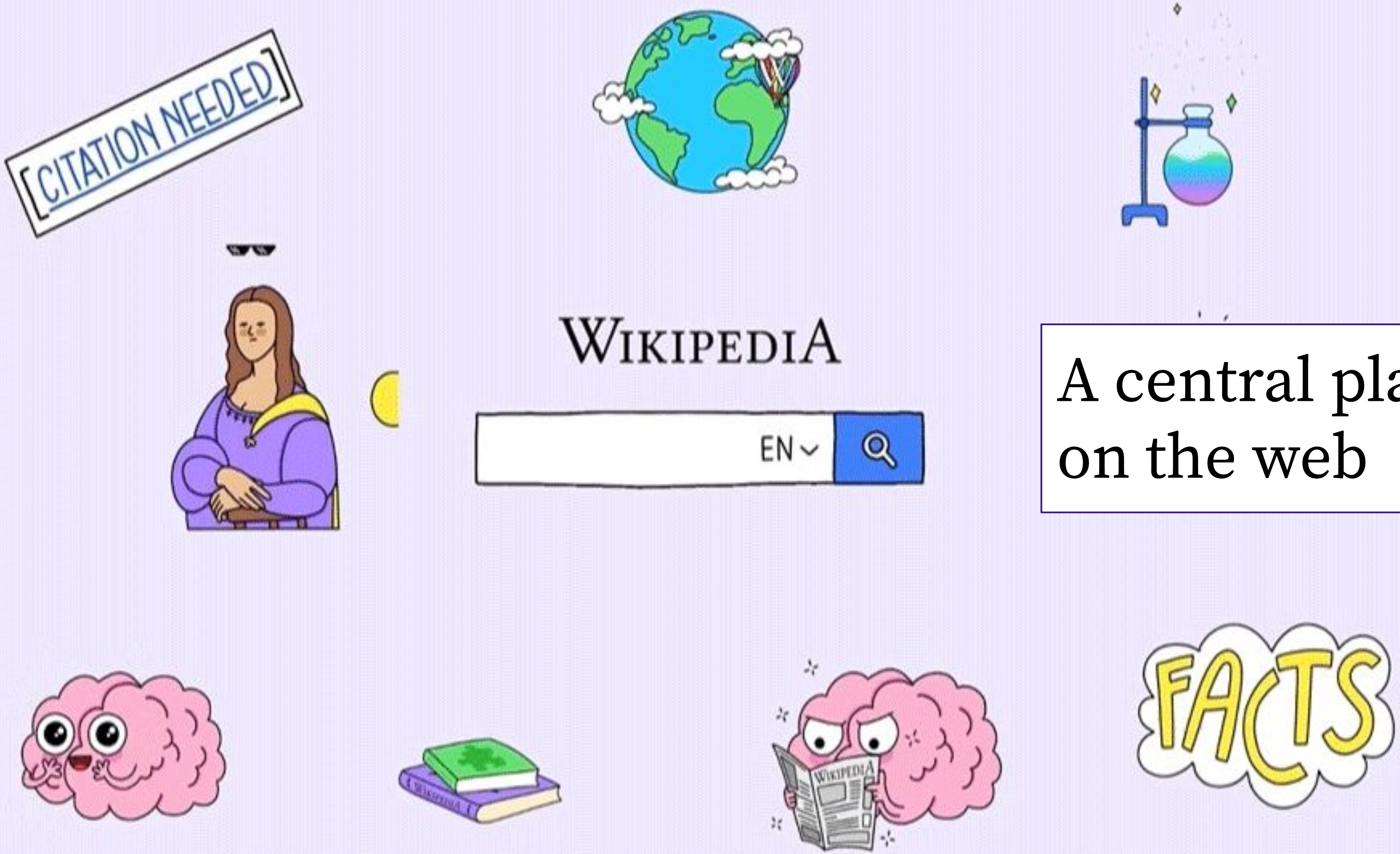
Research ♡ Communities

The Wikimedia Foundation's Research Team



A central place on the web

Research (Tech Dept)

Pablo Aragón

Martin Gerlach

Isaac Johnson

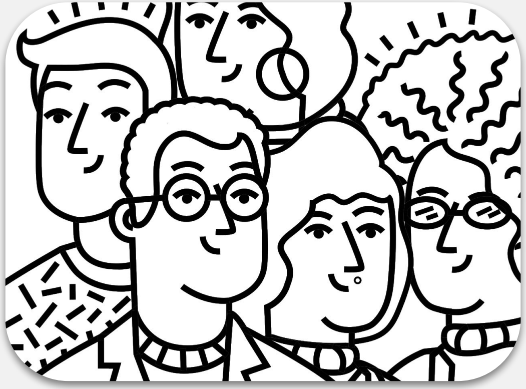
Fabian Kaelin

Emily Lescak

Miriam Redi

Diego Sáez-Trumper

Leila Zia



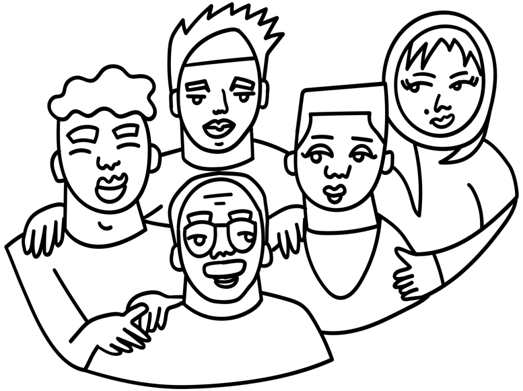
3 contractors, 1 Research Fellow

17 Formal Collaborators

Our Programs

Address Knowledge Gaps

Improve Knowledge Integrity



Grow the Research Community

We envision a world in which **every researcher** can effectively and joyfully **contribute to** the **Wikimedia** projects.

To achieve this vision, we are building a **community of practice** centered on **diversity, inclusiveness, and collaboration**.

Supporting the research community

We organize recurring events

Annual workshops (Wiki Workshop)

Monthly Research Showcases

We award research of potential critical impact for the Wikimedia projects

WMF Research Award of the Year

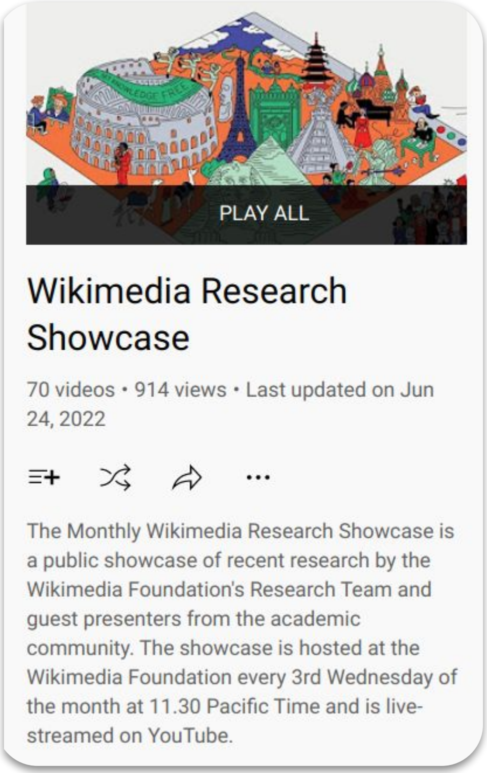
We offer Research Funds

To increase the diversity of the Wikimedia research community

And accelerate research on strategic areas

https://research.wikimedia.org

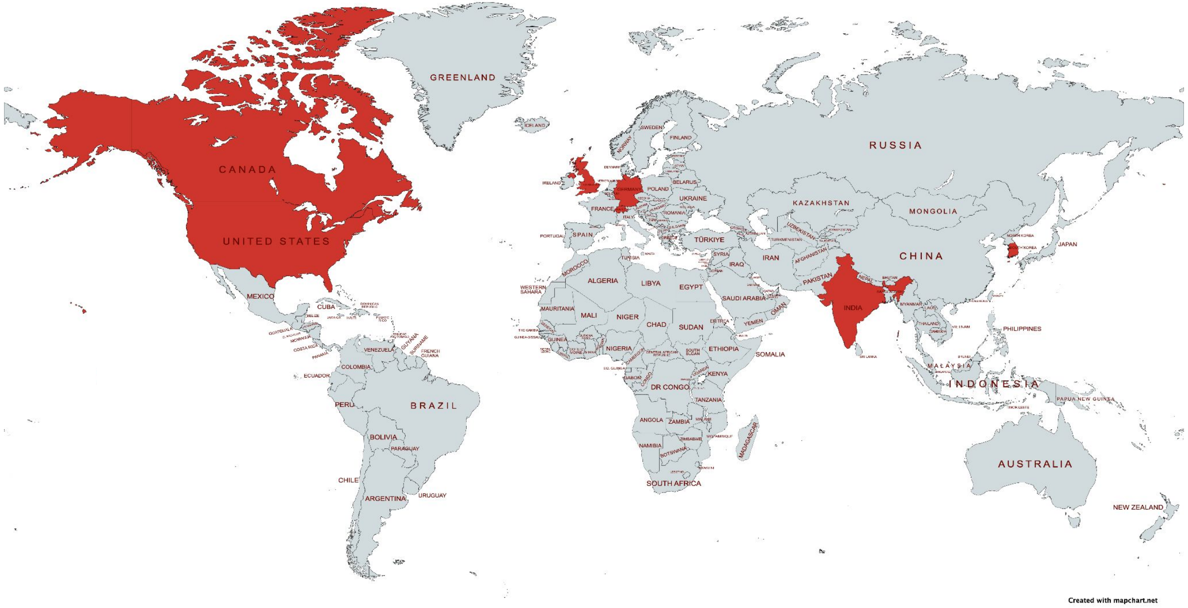
Growing the Research Community ♡



Wikimedia Research Showcase

70 videos • 914 views • Last updated on Jun 24, 2022

The Monthly Wikimedia Research Showcase is a public showcase of recent research by the Wikimedia Foundation's Research Team and guest presenters from the academic community. The showcase is hosted at the Wikimedia Foundation every 3rd Wednesday of the month at 11.30 Pacific Time and is live-streamed on YouTube.



8 showcases in 2022, featuring 20 speakers from 7 countries

Wiki Workshop 2022

Native language

80%

of the workshop attendees report **a language other than English** as their native language. We speak **49 different languages** as our mother tongues.

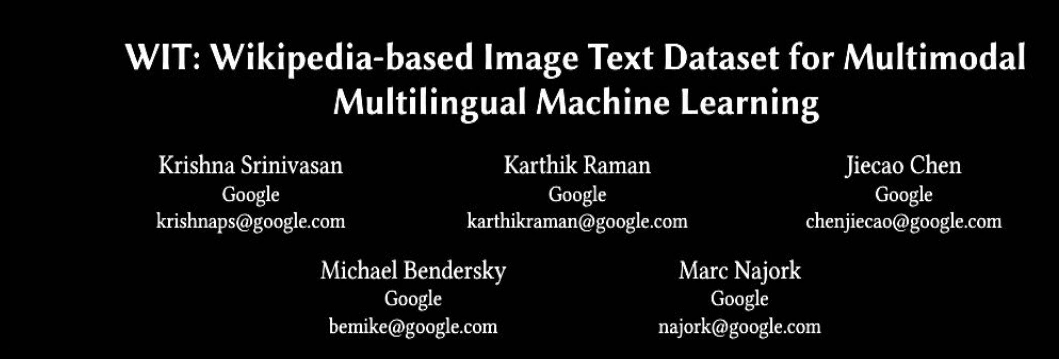
Newcomers

62%

Students

31%

WMF Research Award of the Year 2022



WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning

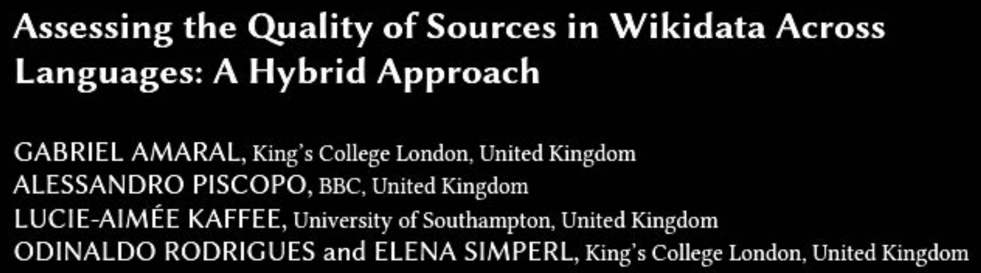
Krishna Srinivasan
Google
krishnaps@google.com

Karthik Raman
Google
karthikraman@google.com

Jiecao Chen
Google
chenjiecao@google.com

Michael Bendersky
Google
bemike@google.com

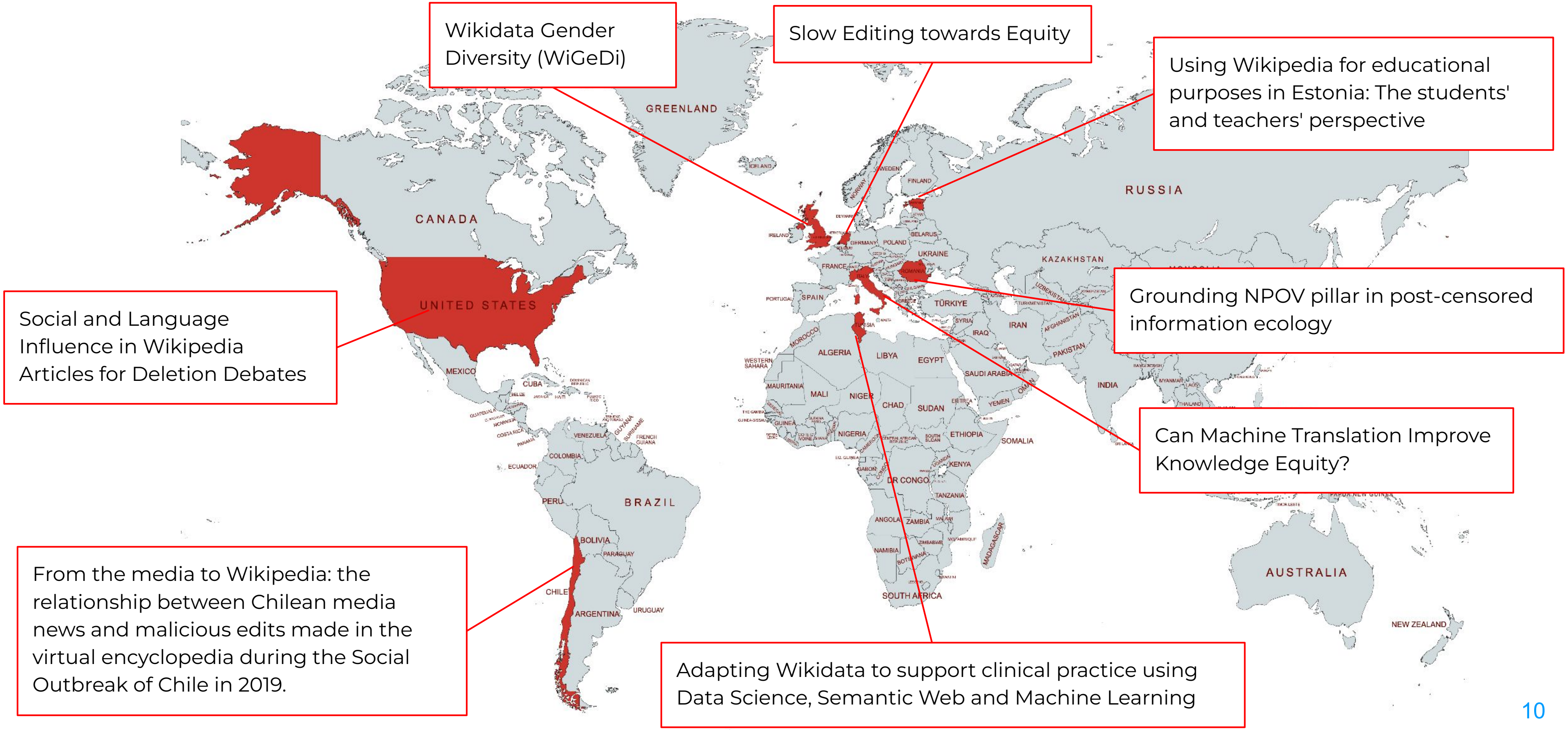
Marc Najork
Google
najork@google.com



Assessing the Quality of Sources in Wikidata Across Languages: A Hybrid Approach

GABRIEL AMARAL, King's College London, United Kingdom
ALESSANDRO PISCOPO, BBC, United Kingdom
LUCIE-AIMÉE KAFFEE, University of Southampton, United Kingdom
ODINALDO RODRIGUES and ELENA SIMPERL, King's College London, United Kingdom

Wikimedia Research Fund 2022



Wikidata Gender Diversity (WiGeDi)

Slow Editing towards Equity

Using Wikipedia for educational purposes in Estonia: The students' and teachers' perspective

Social and Language Influence in Wikipedia Articles for Deletion Debates

Grounding NPOV pillar in post-censored information ecology

Can Machine Translation Improve Knowledge Equity?

From the media to Wikipedia: the relationship between Chilean media news and malicious edits made in the virtual encyclopedia during the Social Outbreak of Chile in 2019.

Adapting Wikidata to support clinical practice using Data Science, Semantic Web and Machine Learning

Created with mapchart.net

Wikimedia Research Fund 2023

Wikimedia Research Fund

Wikimedia Research Fund provides support to individuals, groups, and organizations with research interests on or about Wikimedia projects. We encourage submissions from across research disciplines including but not limited to humanities, social sciences, computer science, education, and law. We aim to support applicants who have limited access to research funding and are proposing work that has potential for direct, positive impact on their local communities.

Want to get involved?

Submit a proposal by December 16.

Comment on proposals when they are posted on Meta-Wiki in January.

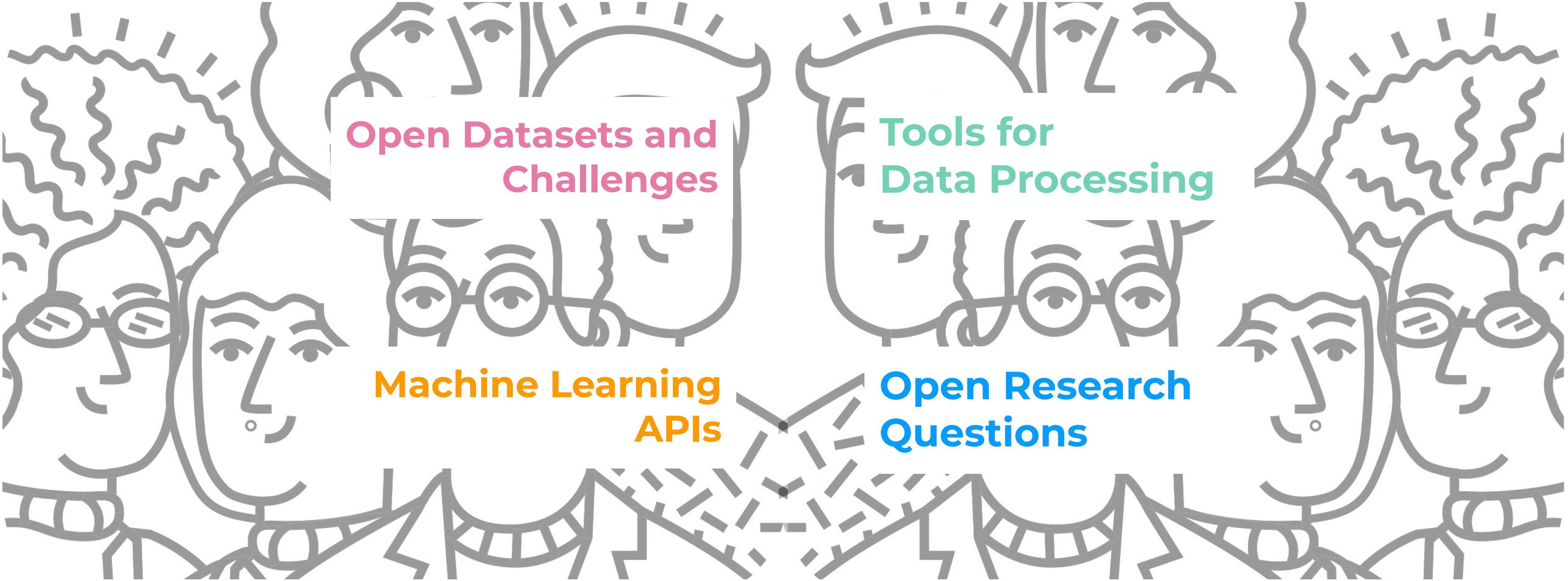
https://meta.wikimedia.org/wiki/Grants:Programs/Wikimedia_Research_%26_Technology_Fund/Wikimedia_Research_Fund

A sustainable distributed network of Wikimedia projects relies on an empowered global community of Wikimedia researchers…

The 4th Wiki Workshop 2017 (Perth)

Image credit: Ricardo Baeza-Yates

… As well as tools and data to encourage a diversity of projects and research questions.



Open Datasets and Challenges

Tools for Data Processing

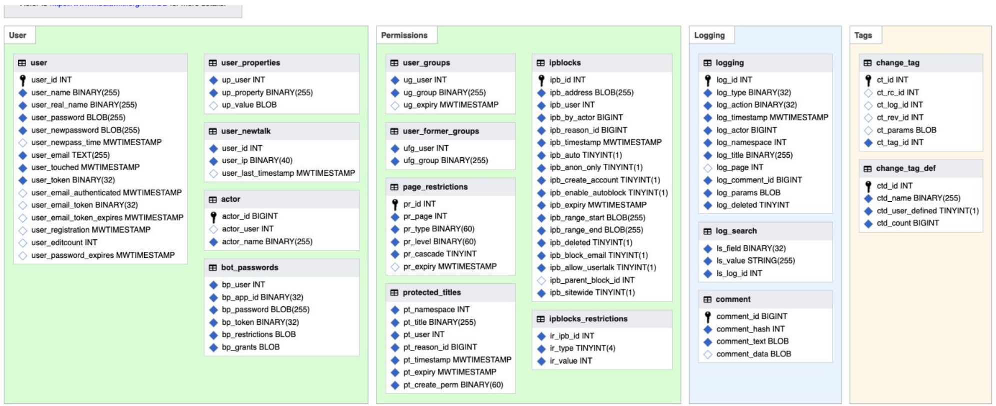
Machine Learning APIs

Open Research Questions

Isaac:

Open Datasets and Challenges

Wikimedia has a lot of data...



User

user
- user_id INT
- user_name BINARY(255)
- user_real_name BINARY(255)
- user_password BLOB(255)
- user_newpassword BLOB(255)
- user_newpass_time MWTIMESTAMP
- user_email TEXT(255)
- user_touched MWTIMESTAMP
- user_token BINARY(32)
- user_email_authenticated MWTIMESTAMP
- user_email_token BINARY(32)
- user_email_token_expires MWTIMESTAMP
- user_registration MWTIMESTAMP
- user_editcount INT
- user_password_expires MWTIMESTAMP

user_properties
- up_user INT
- up_property BINARY(255)
- up_value BLOB

user_newtalk
- user_id INT
- user_ip BINARY(40)
- user_last_timestamp MWTIMESTAMP

actor
- actor_id BIGINT
- actor_user INT
- actor_name BINARY(255)

bot_passwords
- bp_user INT
- bp_app_id BINARY(32)
- bp_password BLOB(255)
- bp_token BINARY(32)
- bp_restrictions BLOB
- bp_grants BLOB

Permissions

user_groups
- ug_user INT
- ug_group BINARY(255)
- ug_expiry MWTIMESTAMP

user_former_groups
- ufg_user INT
- ufg_group BINARY(255)

page_restrictions
- pr_id INT
- pr_page INT
- pr_type BINARY(60)
- pr_level BINARY(60)
- pr_cascade TINYINT
- pr_expiry MWTIMESTAMP

protected_titles
- pt_namespace INT
- pt_title BINARY(255)
- pt_user INT
- pt_reason_id BIGINT
- pt_timestamp MWTIMESTAMP
- pt_expiry MWTIMESTAMP
- pt_create_perm BINARY(60)

ipblocks
- ipb_id INT
- ipb_address BLOB(255)
- ipb_user INT
- ipb_by_actor BIGINT
- ipb_reason_id BIGINT
- ipb_timestamp MWTIMESTAMP
- ipb_auto TINYINT(1)
- ipb_anon_only TINYINT(1)
- ipb_create_account TINYINT(1)
- ipb_enable_autoblock TINYINT(1)
- ipb_expiry MWTIMESTAMP
- ipb_range_start BLOB(255)
- ipb_range_end BLOB(255)
- ipb_deleted TINYINT(1)
- ipb_block_email TINYINT(1)
- ipb_allow_usertalk TINYINT(1)
- ipb_parent_block_id INT
- ipb_sitewide TINYINT(1)

ipblocks_restrictions
- ir_ipb_id INT
- ir_type TINYINT(4)
- ir_value INT

Logging

logging
- log_id INT
- log_type BINARY(32)
- log_action BINARY(32)
- log_timestamp MWTIMESTAMP
- log_actor BIGINT
- log_namespace INT
- log_title BINARY(255)
- log_page INT
- log_comment_id BIGINT
- log_params BLOB
- log_deleted TINYINT

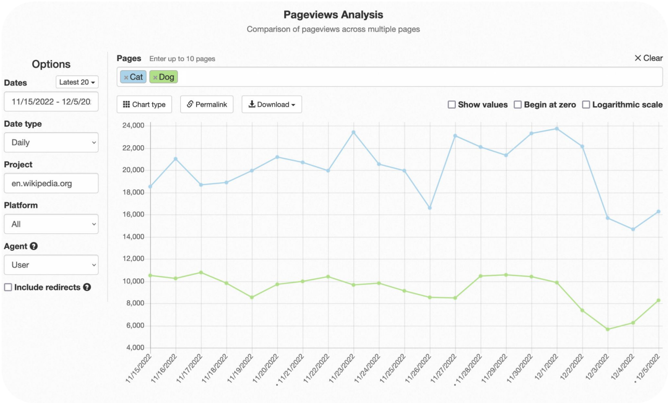
log_search
- ls_field BINARY(32)
- ls_value STRING(255)
- ls_log_id INT

comment
- comment_id BIGINT
- comment_hash INT
- comment_text BLOB
- comment_data BLOB

Tags

change_tag
- ct_id INT
- ct_rc_id INT
- ct_log_id INT
- ct_rev_id INT
- ct_params BLOB
- ct_tag_id INT

change_tag_def
- ctd_id INT
- ctd_name BINARY(255)
- ctd_user_defined TINYINT(1)
- ctd_count BIGINT



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

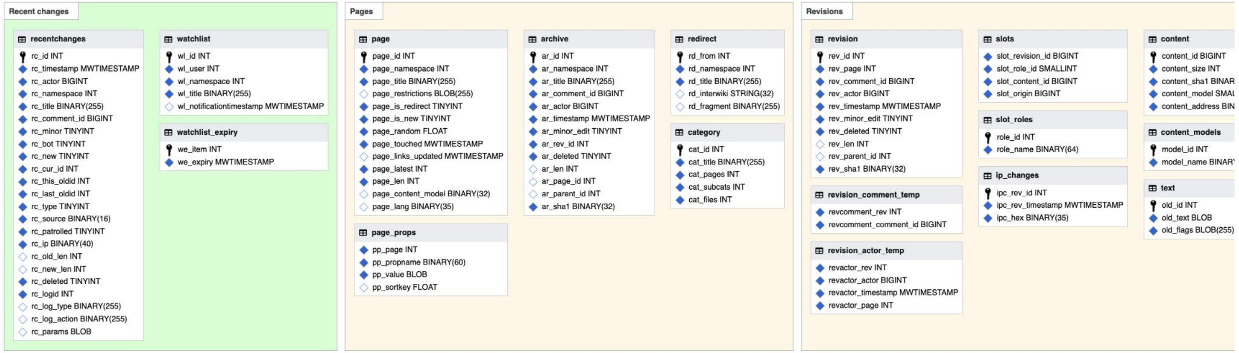
Agent: User

□ Include redirects

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

Chart type | Permalink | Download ▾

□ Show values □ Begin at zero □ Logarithmic scale



Recent changes

recentchanges
- rc_id INT
- rc_timestamp MWTIMESTAMP
- rc_actor BIGINT
- rc_namespace INT
- rc_title BINARY(255)
- rc_comment_id BIGINT
- rc_minor TINYINT
- rc_bot TINYINT
- rc_new TINYINT
- rc_cur_id INT
- rc_this_oldid INT
- rc_last_oldid INT
- rc_type TINYINT
- rc_source BINARY(16)
- rc_patrolled TINYINT
- rc_ip BINARY(40)
- rc_old_len INT
- rc_new_len INT
- rc_deleted TINYINT
- rc_logid INT
- rc_log_type BINARY(255)
- rc_log_action BINARY(255)
- rc_params BLOB

watchlist
- wl_id INT
- wl_user INT
- wl_namespace INT
- wl_title BINARY(255)
- wl_notificationtimestamp MWTIMESTAMP

watchlist_expiry
- we_item INT
- we_expiry MWTIMESTAMP

Pages

page
- page_id INT
- page_namespace INT
- page_title BINARY(255)
- page_restrictions BLOB(255)
- page_is_redirect TINYINT
- page_is_new TINYINT
- page_random FLOAT
- page_touched MWTIMESTAMP
- page_links_updated MWTIMESTAMP
- page_latest INT
- page_len INT
- page_content_model BINARY(32)
- page_lang BINARY(35)

page_props
- pp_page INT
- pp_propname BINARY(60)
- pp_value BLOB
- pp_sortkey FLOAT

archive
- ar_id INT
- ar_namespace INT
- ar_title BINARY(255)
- ar_comment_id BIGINT
- ar_actor BIGINT
- ar_timestamp MWTIMESTAMP
- ar_minor_edit TINYINT
- ar_rev_id INT
- ar_deleted TINYINT
- ar_len INT
- ar_page_id INT
- ar_parent_id INT
- ar_sha1 BINARY(32)

redirect
- rd_from INT
- rd_namespace INT
- rd_title BINARY(255)
- rd_interwiki STRING(32)
- rd_fragment BINARY(255)

category
- cat_id INT
- cat_title BINARY(255)
- cat_pages INT
- cat_subcats INT
- cat_files INT

Revisions

revision
- rev_id INT
- rev_page INT
- rev_comment_id BIGINT
- rev_actor BIGINT
- rev_timestamp MWTIMESTAMP
- rev_minor_edit TINYINT
- rev_deleted TINYINT
- rev_len INT
- rev_parent_id INT
- rev_sha1 BINARY(32)

revision_comment_temp
- revcomment_rev INT
- revcomment_comment_id BIGINT

revision_actor_temp
- revactor_rev INT
- revactor_actor BIGINT
- revactor_timestamp MWTIMESTAMP
- revactor_page INT

slots
- slot_revision_id BIGINT
- slot_role_id SMALLINT
- slot_content_id BIGINT
- slot_origin BIGINT

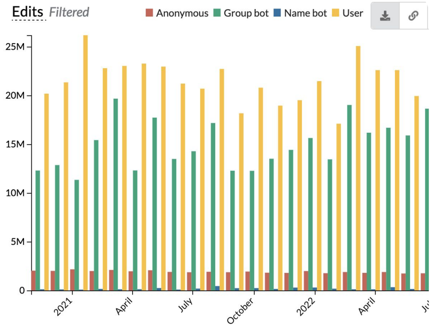
slot_roles
- role_id INT
- role_name BINARY(64)

ip_changes
- ipc_rev_id INT
- ipc_rev_timestamp MWTIMESTAMP
- ipc_hex BINARY(35)

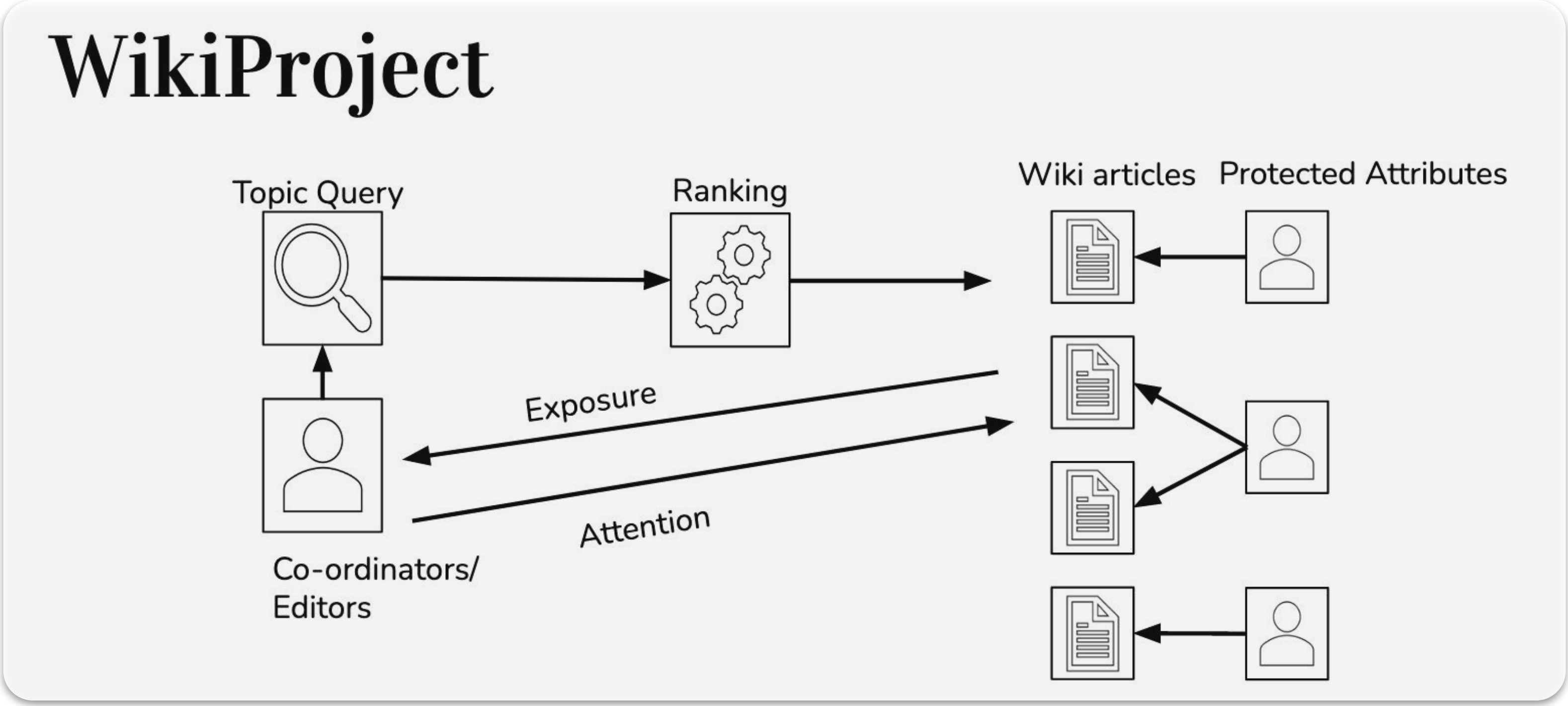
content
- content_id BIGINT
- content_size INT
- content_sha1 BINAR
- content_model SMAL
- content_address BIN

content_models
- model_id INT
- model_name BINAR

text
- old_id INT
- old_text BLOB
- old_flags BLOB(255)

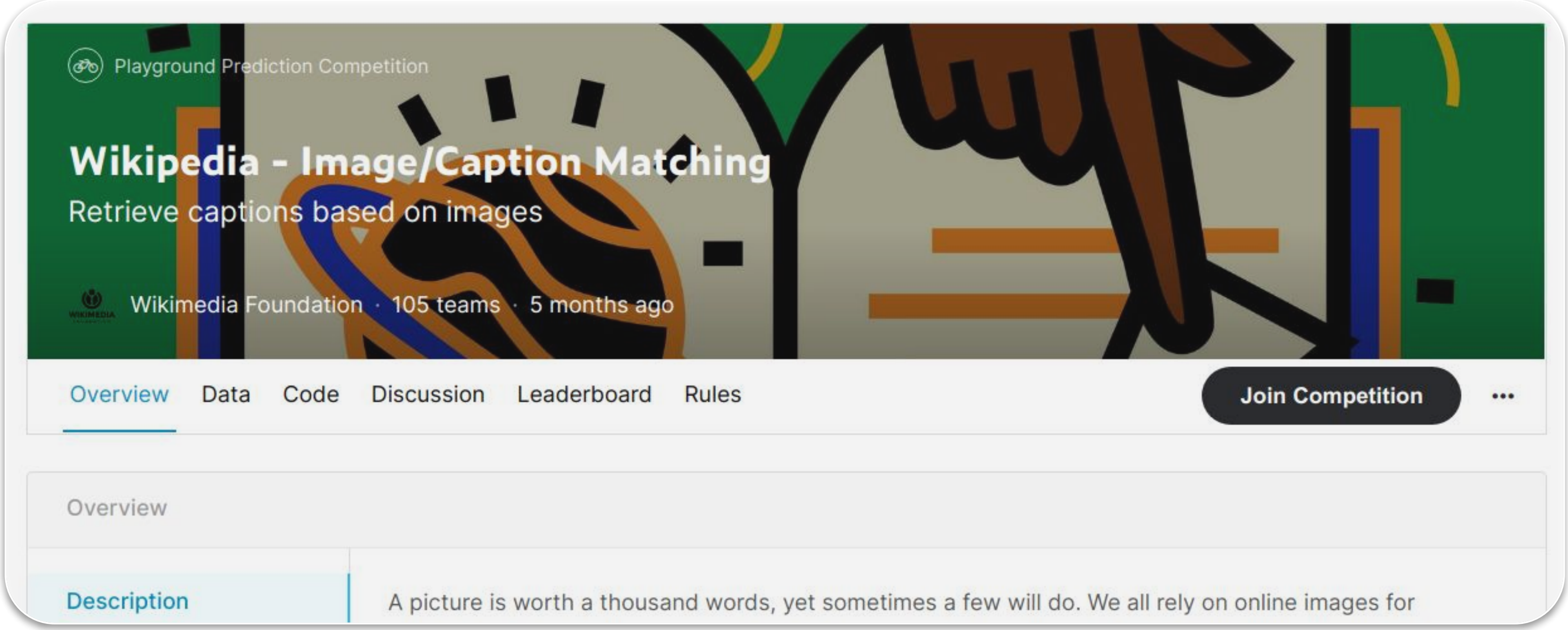


TREC Fair Ranking (2021-22)



9 teams + papers describing each approach:
https://fair-trec.github.io/

Wikipedia Image/Caption Competition



Playground Prediction Competition

Wikipedia - Image/Caption Matching

Retrieve captions based on images

Wikimedia Foundation · 105 teams · 5 months ago

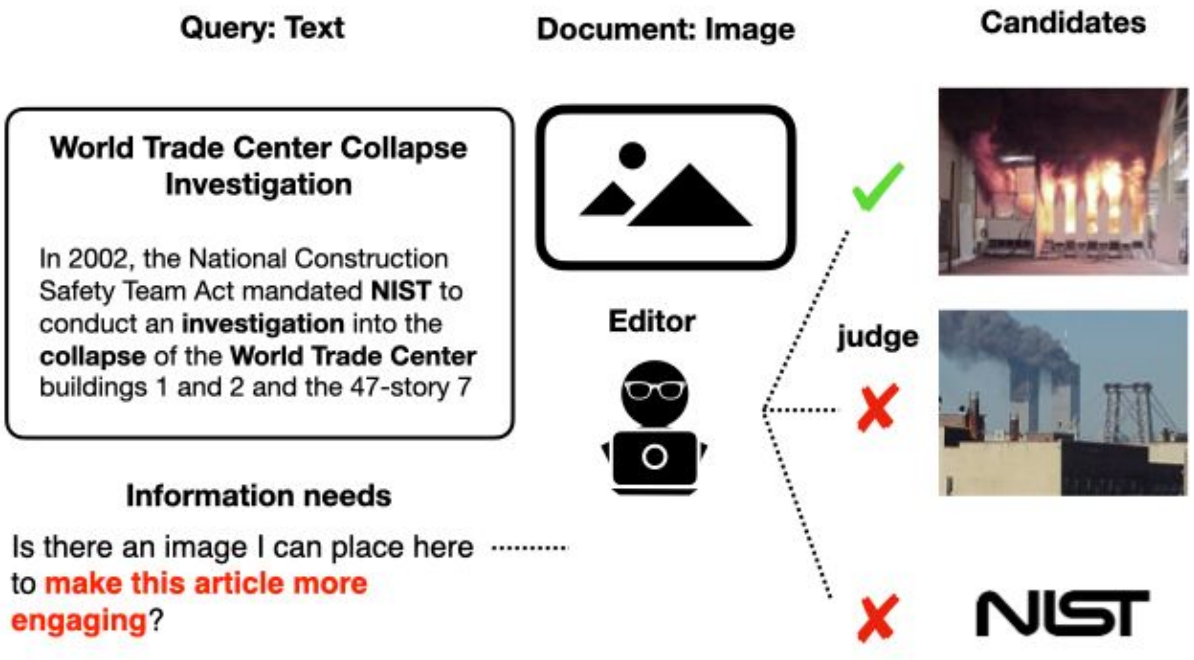
Overview

Description

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

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

TREC Multimedia (AToMiC)



Query: Text

Document: Image

Candidates

World Trade Center Collapse Investigation

In 2002, the National Construction Safety Team Act mandated **NIST** to conduct an **investigation** into the **collapse** of the **World Trade Center** buildings 1 and 2 and the 47-story 7

Information needs

Is there an image I can place here to **make this article more engaging**?

Editor

judge

✓

✗

✗

NIST

Learn more: https://trec-atomic.github.io/about/

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:

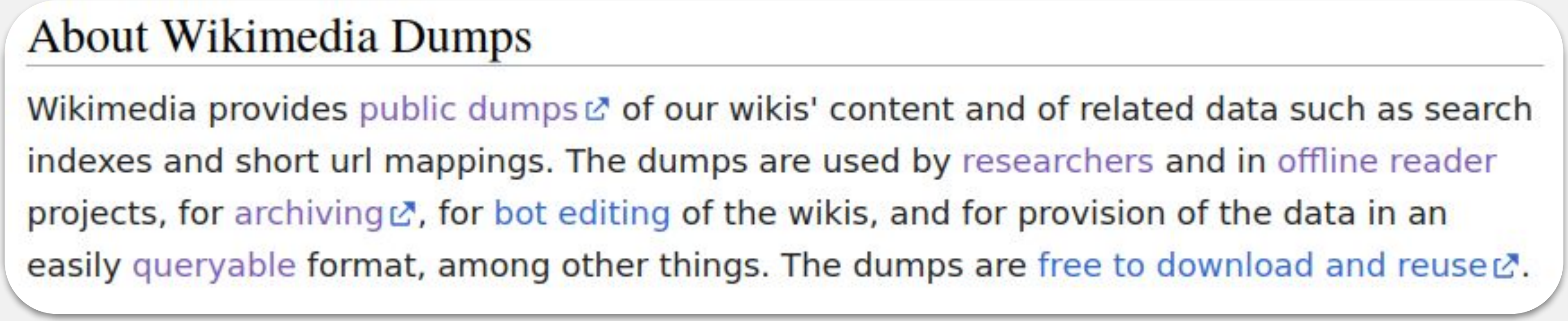
https://w.wiki/64CE

-

Martin:

Data Analysis and Processing Tools

Making the Data Available

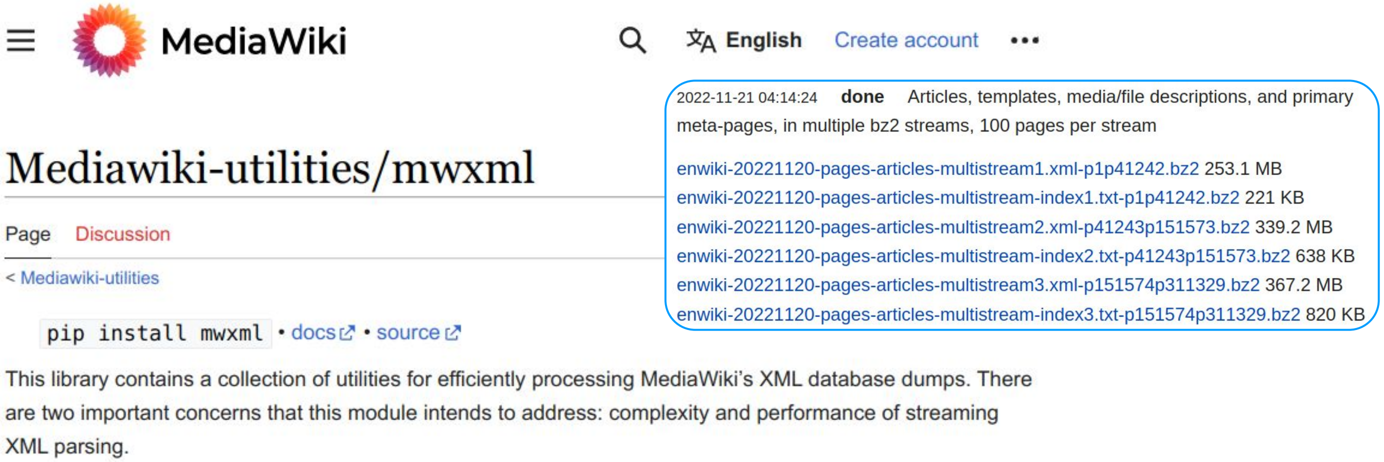


About Wikimedia Dumps

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

Making the Data ~~Available~~ Accessible



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

Mediawiki-utilities/mwxml

Page Discussion

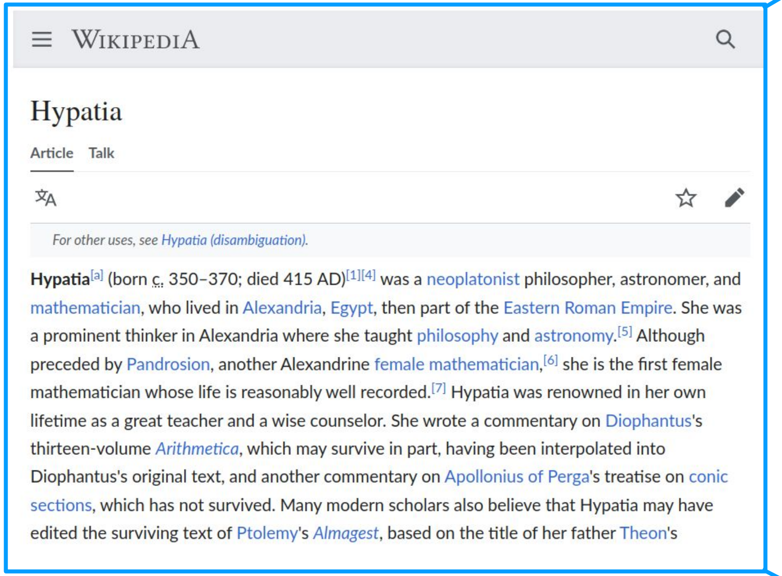
< Mediawiki-utilities

`pip install mwxml` • docs • source

This library contains a collection of utilities for efficiently processing MediaWiki's XML database dumps. There are two important concerns that this module intends to address: complexity and performance of streaming XML parsing.

mwxml: streaming xml-files

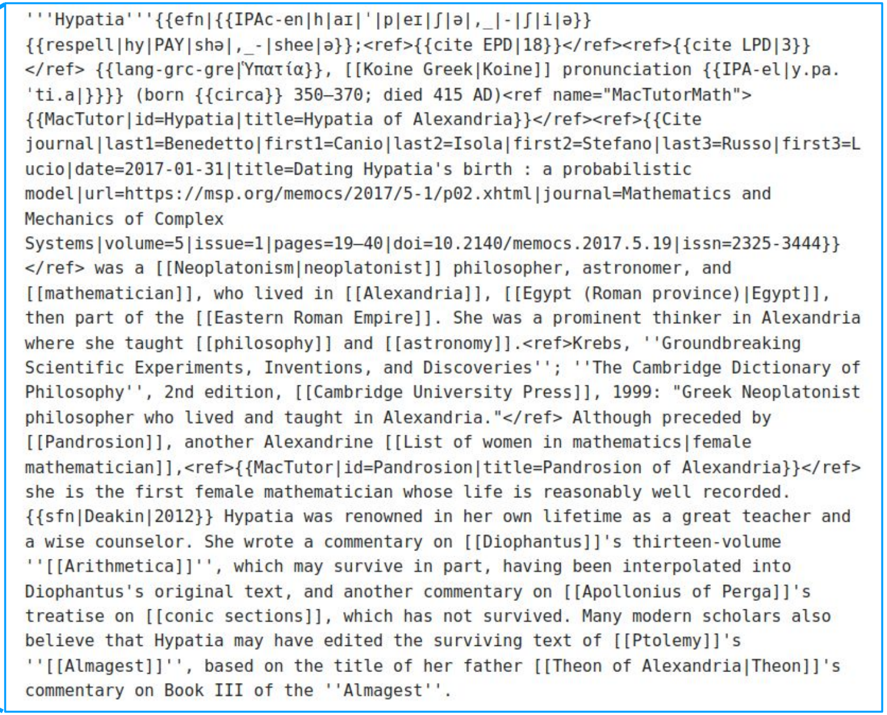
Making the Data ~~Available~~ Accessible



Hypatia

For other uses, see Hypatia (disambiguation).

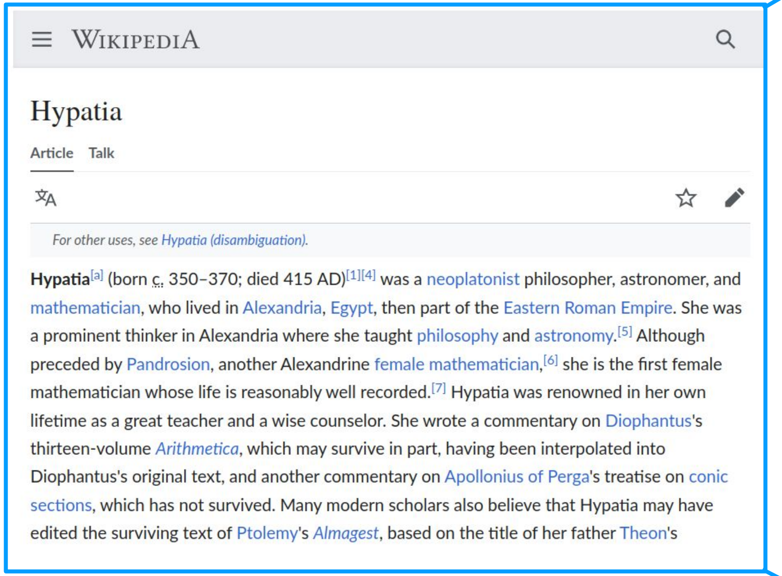
Hypatia[a] (born c. 350–370; died 415 AD)[1][4] was a neoplatonist philosopher, astronomer, and mathematician, who lived in Alexandria, Egypt, then part of the Eastern Roman Empire. She was a prominent thinker in Alexandria where she taught philosophy and astronomy.[5] Although preceded by Pandrosion, 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



```
'''Hypatia'''{{efn|{{IPAc-en|h|aɪ|'|p|eɪ|ʃ|ə|,_|-|ʃ|i|ə}}
{{respell|hy|PAY|shə|,_-|shee|ə}};<ref>{{cite EPD|18}}</ref><ref>{{cite LPD|3}}
</ref> {{lang-grc-gre|Ὑπατία}}, [[Koine Greek|Koine]] pronunciation {{IPA-el|y.pa.
'ti.a|}}}} (born {{circa}} 350–370; died 415 AD)<ref name="MacTutorMath">
{{MacTutor|id=Hypatia|title=Hypatia of Alexandria}}</ref><ref>{{Cite
journal|last1=Benedetto|first1=Canio|last2=Isola|first2=Stefano|last3=Russo|first3=L
ucio|date=2017-01-31|title=Dating Hypatia's birth : a probabilistic
model|url=https://msp.org/memocs/2017/5-1/p02.xhtml|journal=Mathematics and
Mechanics of Complex
Systems|volume=5|issue=1|pages=19–40|doi=10.2140/memocs.2017.5.19|issn=2325-3444}}
</ref> was a [[Neoplatonism|neoplatonist]] philosopher, astronomer, and
[[mathematician]], who lived in [[Alexandria]], [[Egypt (Roman province)|Egypt]],
then part of the [[Eastern Roman Empire]]. She was a prominent thinker in Alexandria
where she taught [[philosophy]] and [[astronomy]].<ref>Krebs, ''Groundbreaking
Scientific Experiments, Inventions, and Discoveries''; ''The Cambridge Dictionary of
Philosophy'', 2nd edition, [[Cambridge University Press]], 1999: "Greek Neoplatonist
philosopher who lived and taught in Alexandria."</ref> Although preceded by
[[Pandrosion]], another Alexandrine [[List of women in mathematics|female
mathematician]],<ref>{{MacTutor|id=Pandrosion|title=Pandrosion of Alexandria}}</ref>
she is the first female mathematician whose life is reasonably well recorded.
{{sfn|Deakin|2012}} 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 of Alexandria|Theon]]'s
commentary on Book III of the ''Almagest''.
```

mwparserfromhell: parsing wikitext

Making the Data ~~Available~~ Accessible



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

Links

Templates

References

Plain Text

...

mwparserfromhell: parsing wikitext

Tools we developed

WikiNav (with Outreachy intern Muniza A.)

graphical UI for clickstream data dumps

Mwsql (with Outreachy intern Slawina S.)

efficiently processing MediaWiki's SQL database dumps

Mwedittypes (with Outreachy intern Jesse A.)

structured parsing of edit-diffs

Mwparserfromhtml (with Outreachy intern Nazia T.)

mining and parsing HTML dumps

HTML Dumps

Wikimedia Enterprise
HTML Dumps

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

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

New Dump dataset (Oct 2021)

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

Why are HTML-dumps exciting?

More content

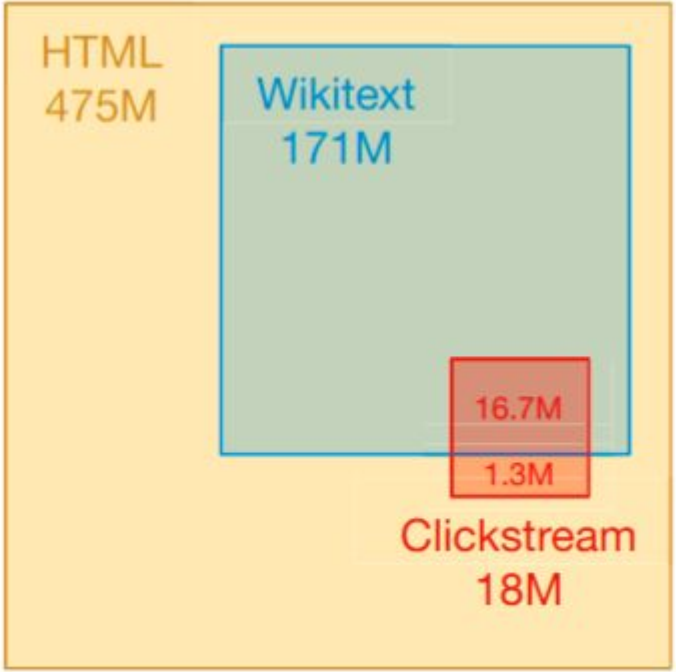
wikitext misses a lot of the article’s content;

HTML-version captures what a reader sees;

More complexity

Building HTML dumps is expensive

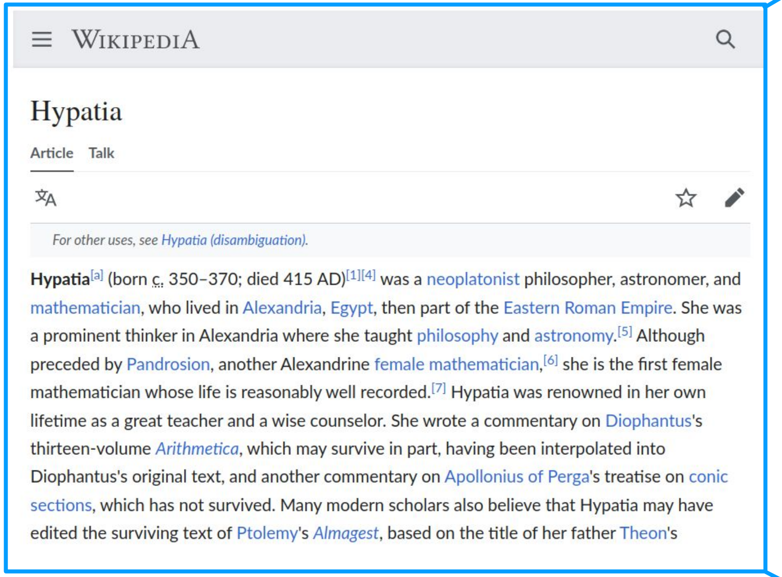
Parsing even a single snapshot of full English Wikipedia from wikitext to HTML via the Wikipedia API takes about 5 days at maximum speed.



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

mwparserfromhtml

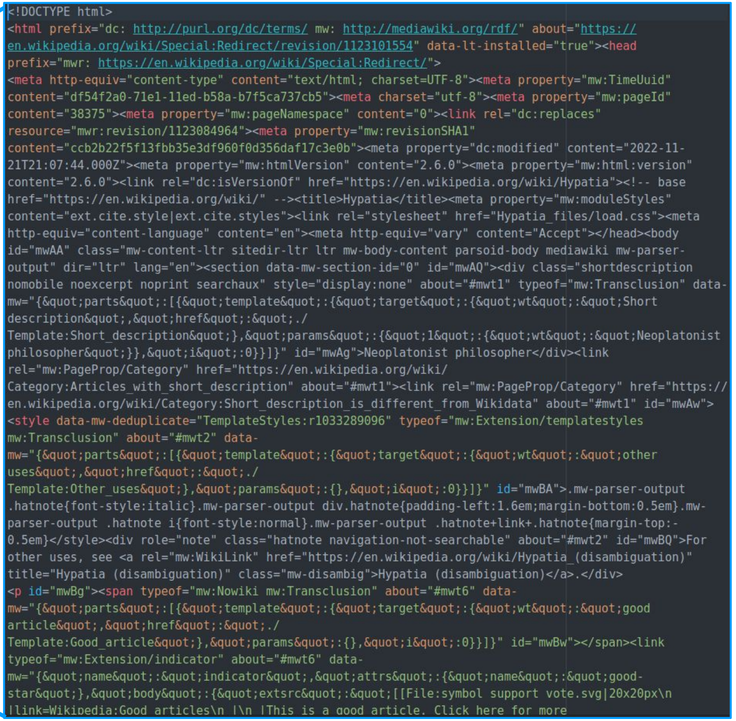
Challenge: How to (easily) parse the HTML of an article?



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

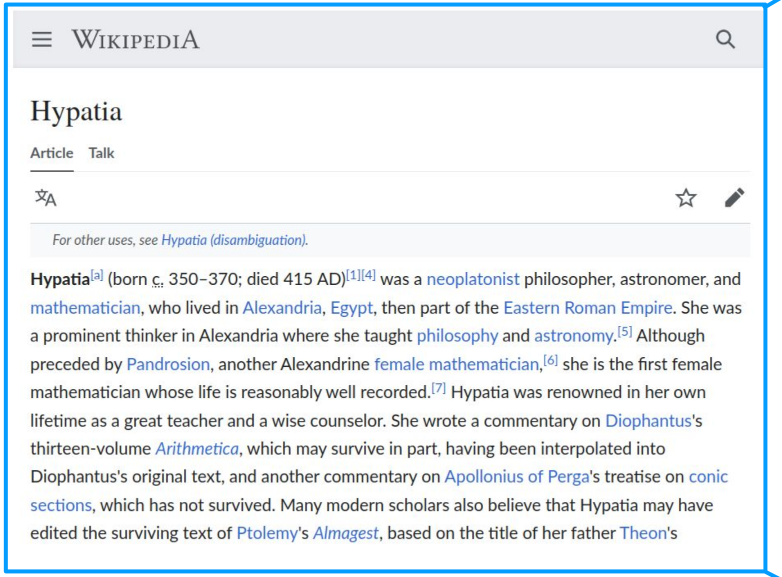


```
<!DOCTYPE html>
<html prefix="dc: http://purl.org/dc/terms/ mw: http://mediawiki.org/rdf/" about="https://
en.wikipedia.org/wiki/Special:Redirect/revision/1123101554" data-lt-installed="true"><head
prefix="mwr: https://en.wikipedia.org/wiki/Special:Redirect/">
<meta http-equiv="content-type" content="text/html; 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

Solution: Extract features from the HTML

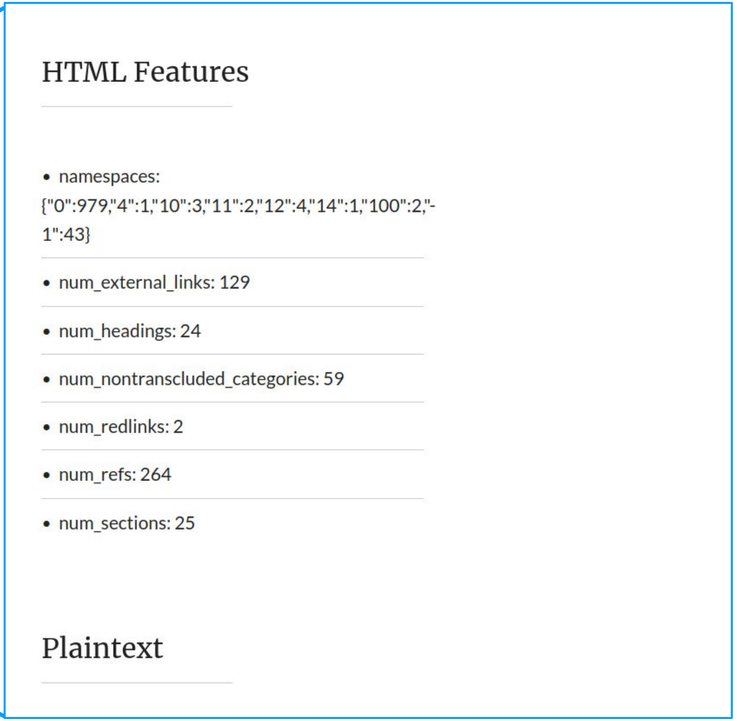


Hypatia

Article Talk

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



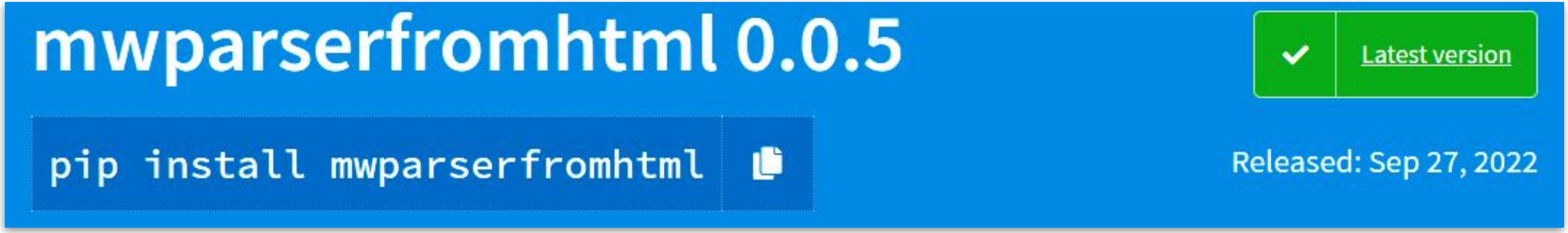
HTML Features

- namespaces: {"0":979,"4":1,"10":3,"11":2,"12":4,"14":1,"100":2,"-1":43}
- num_external_links: 129
- num_headings: 24
- num_nontranscluded_categories: 59
- num_redlinks: 2
- num_refs: 264
- num_sections: 25

Plaintext

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

mwparserfromhtml



mwparserfromhtml 0.0.5

```
pip install mwparserfromhtml
```

Released: Sep 27, 2022

Python library for parsing HTML Dumps

Work in progress. Contributions are welcome

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

Diego:

Machine Learning APIs

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

Experimental vs Production APIs

Experimental APIs

Hosted on ToolForge or WMF's VPS Cloud

Handles small amount of requests.

Our latest (under development) models.

Production APIs

Stable

Hosted by other teams (ML-Platform/Product)

Designed to handle large amount of requests

Experimental vs Production APIs

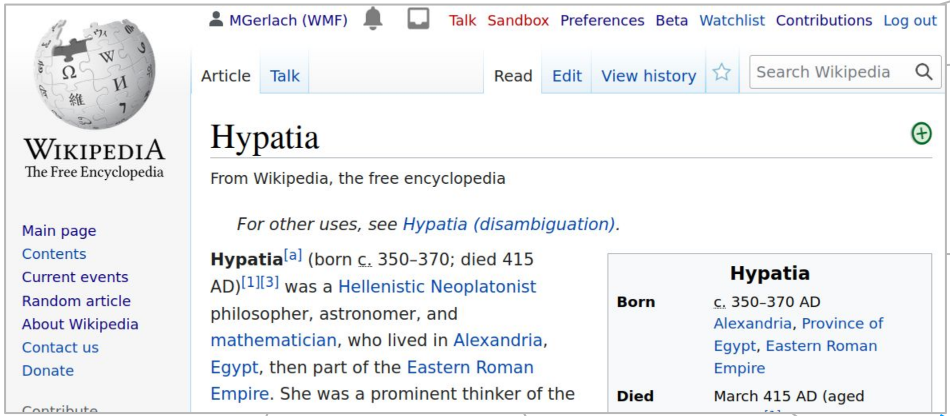
Experimental APIs

- Image similarity model
- Article description generation model
- Article Quality

Production APIs

- Article Topic
- Link Recommendation
- **Revert Risk Model**

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:

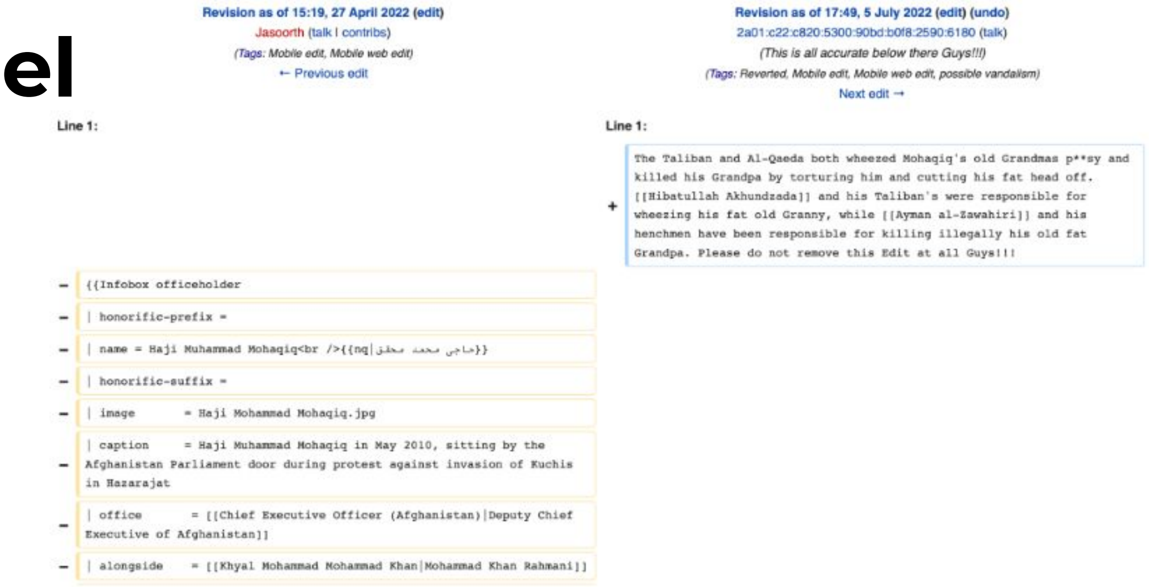
Biography
Philosophy/Religion
STEM

Readability:

Medium

Revert Risk Model

Support patrollers with the identification of revisions that might need to be reverted



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]]
```

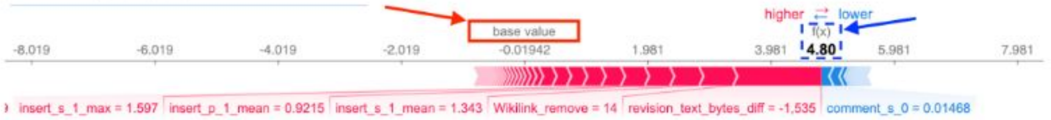
Model performance:

True value: **IS_REVERT**

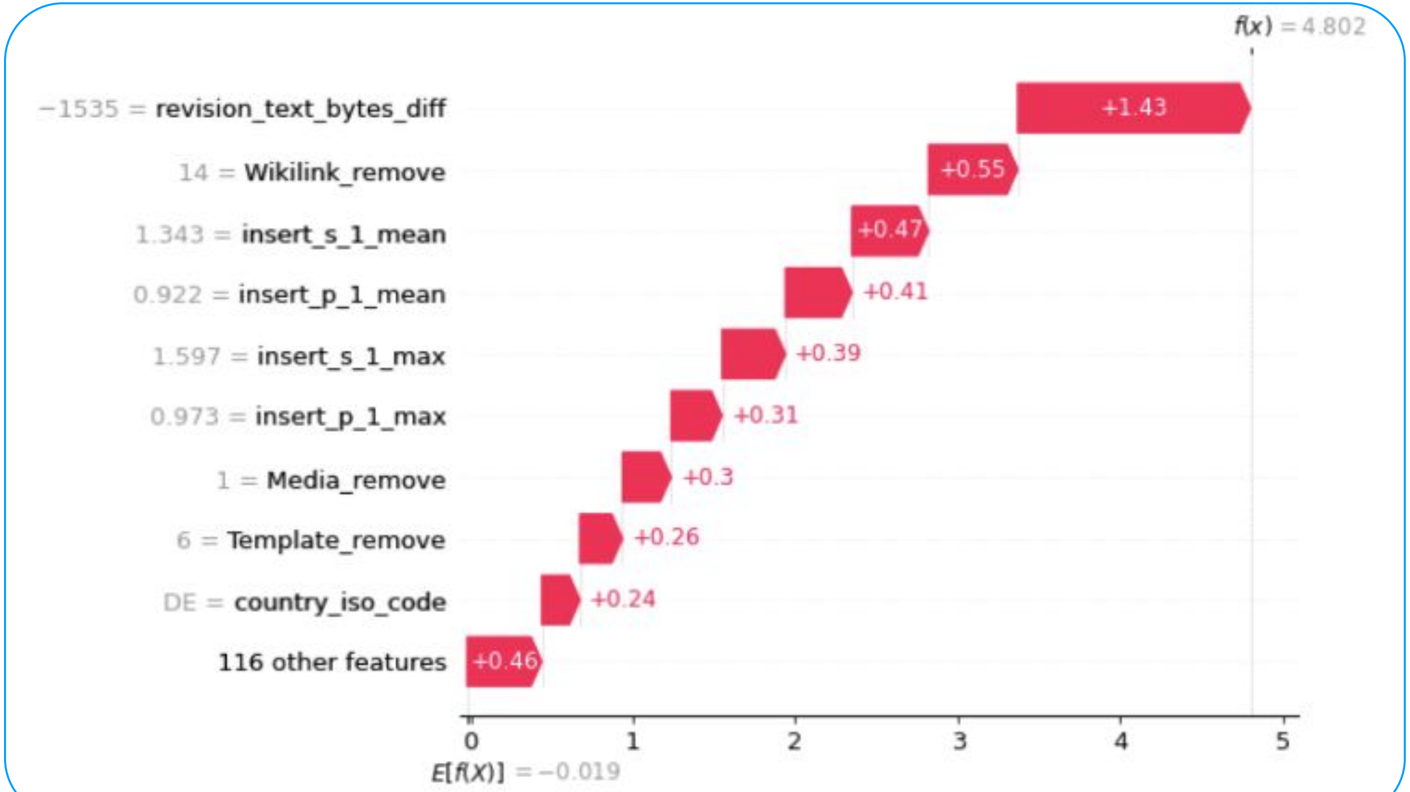
Predicted value: **IS_REVERT**

IS_REVERT predicted probability: 0.992

SHAP Explanation:



Revert Risk Model: explainability



Revert Risk Model: biases/performance

	All Edits	Anonymous Edits
Language Agnostic	0.79	0.67
Multilingual	0.68	0.69

F1 SCORE FOR Revert Risk Model (Balanced data)

Fully language agnostic model

Considers users information (tenure & #revisions). This might introduces biases against new users.

Multilingual Model

Works better for anonymous edits.

Next steps

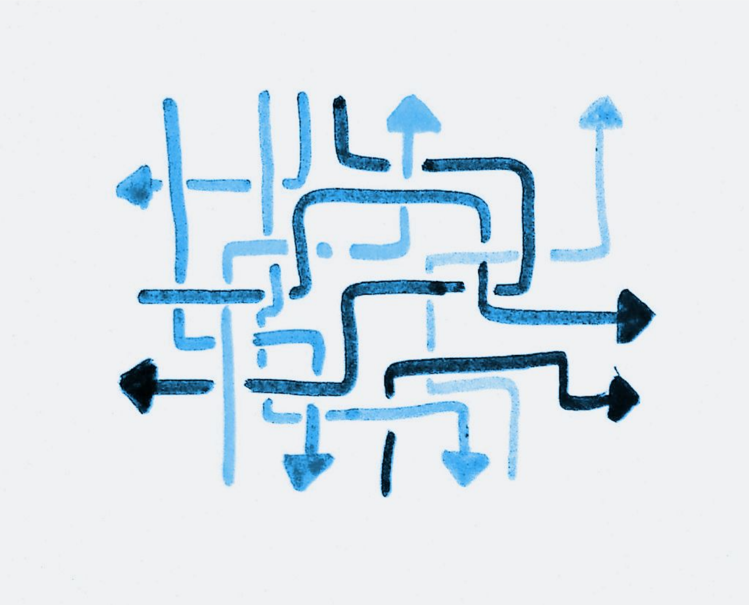
Model Productization

Put multilingual model in production

Public facing API

For Revert Risk language-agnostic model and article quality.

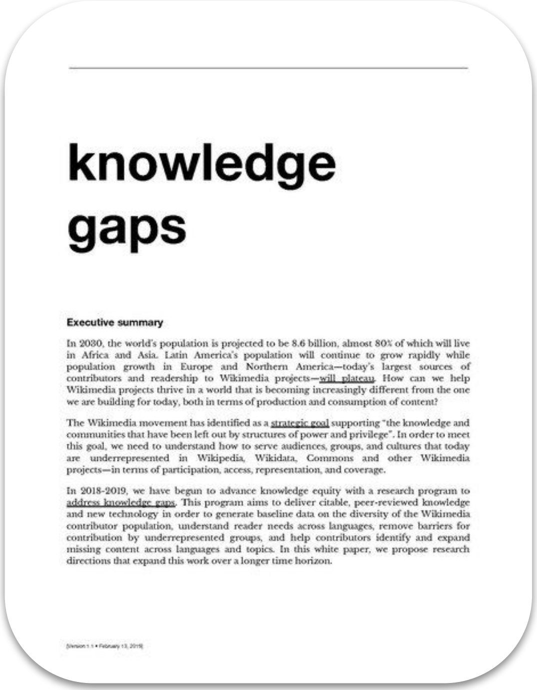
New models for Wikidata



Pablo:

Surfacing Open Research Questions

Recommended research directions



knowledge gaps

Executive summary

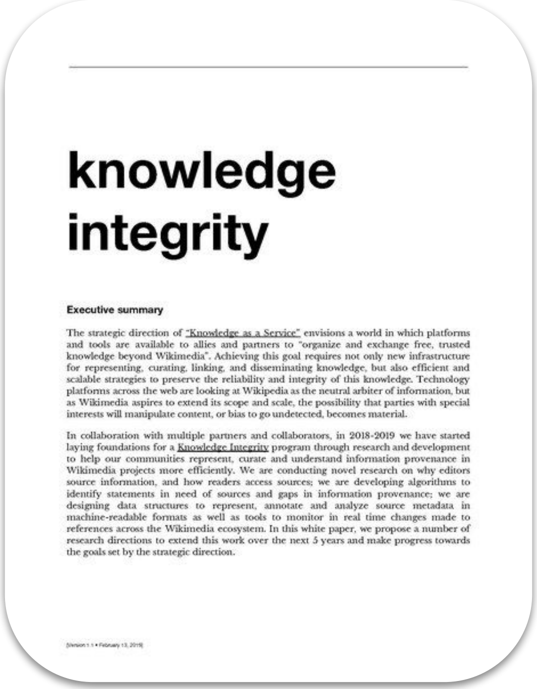
In 2030, the world's population is projected to be 8.6 billion, almost 80% of which will live in Africa and Asia. Latin America's population will continue to grow rapidly while population growth in Europe and Northern America—today's largest sources of contributors and readership to Wikimedia projects—will 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]

doi.org/10.6084/m9.figshare.7698245



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]

doi.org/10.6084/m9.figshare.7704626

foundations

Executive summary

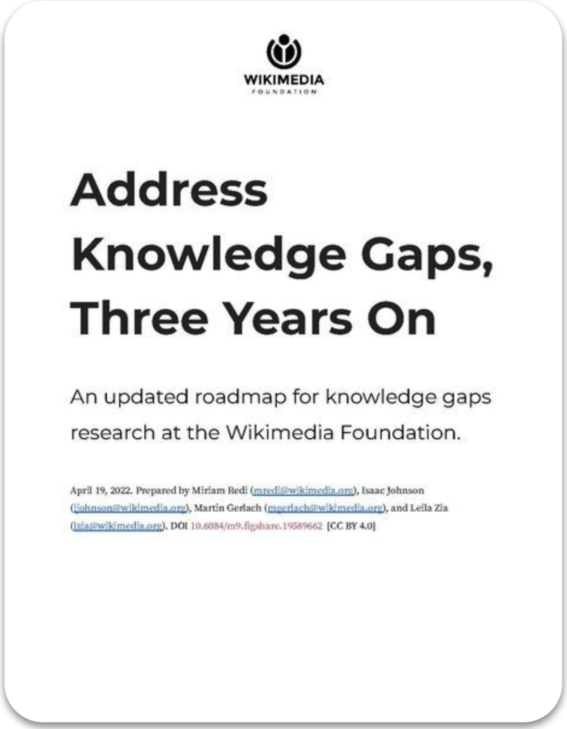
Wikimedia projects are created and maintained by a vast network of individual contributors and organizations with different roles and expertise. The Wikimedia Foundation, including Wikimedia Research, plays an important role in supporting these efforts, but our internal capacity and expertise will always 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]

doi.org/10.6084/m9.figshare.7704629

Knowledge gaps: updated roadmap



WIKIMEDIA FOUNDATION

Address Knowledge Gaps, Three Years On

An updated roadmap for knowledge gaps research at the Wikimedia Foundation.

April 19, 2022. Prepared by Miriam Redi (mredi@wikimedia.org), Isaac Johnson (ijohnson@wikimedia.org), Martin Gerlach (mgerlach@wikimedia.org), and Leila Zia (lzia@wikimedia.org). DOI 10.6084/m9.figshare.19589662 [CC BY 4.0]

Guiding Principles

Consolidated Research Areas

Ideas for Future Research

Learning and Wikipedia

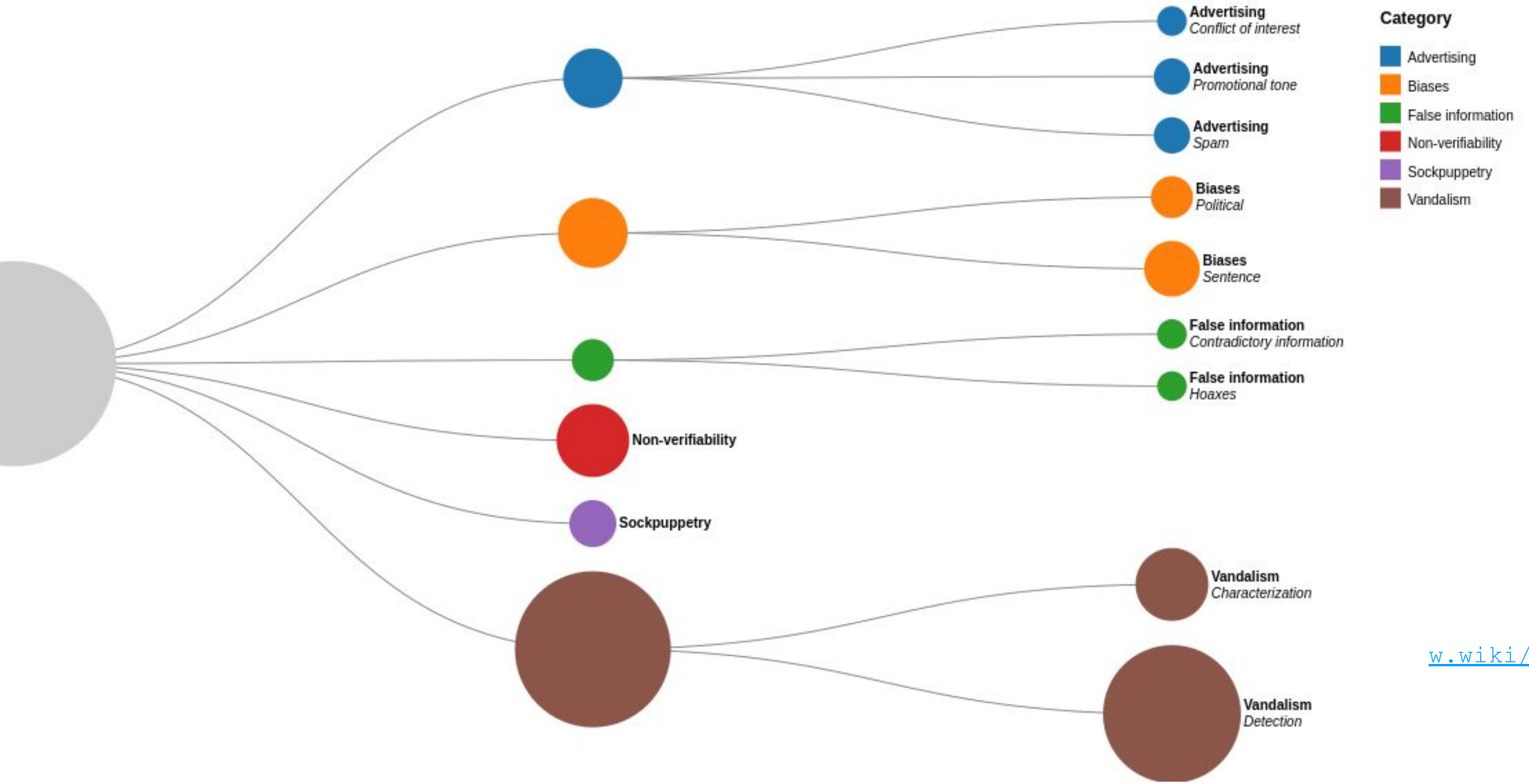
A **Model** of Wikipedia’s **Complexity**

Named **Entity Recognition** in **images**

New and **External** forms of **knowledge**

meta.wikimedia.org/wiki/Research:Knowledge_Gaps_3_Years_On

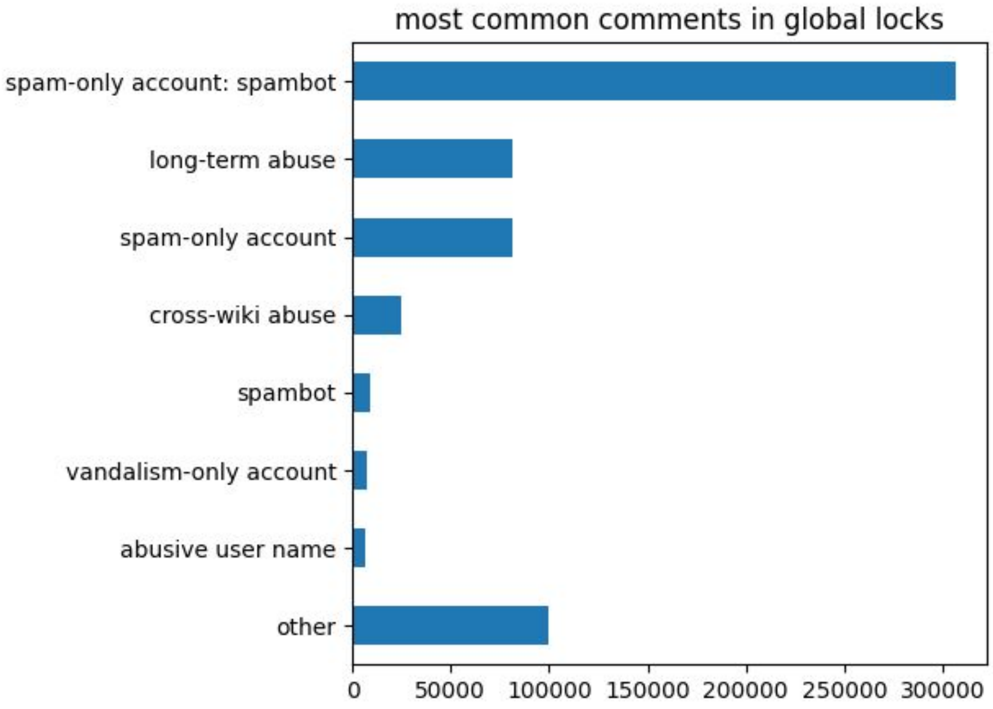
Knowledge Integrity: literature mapping

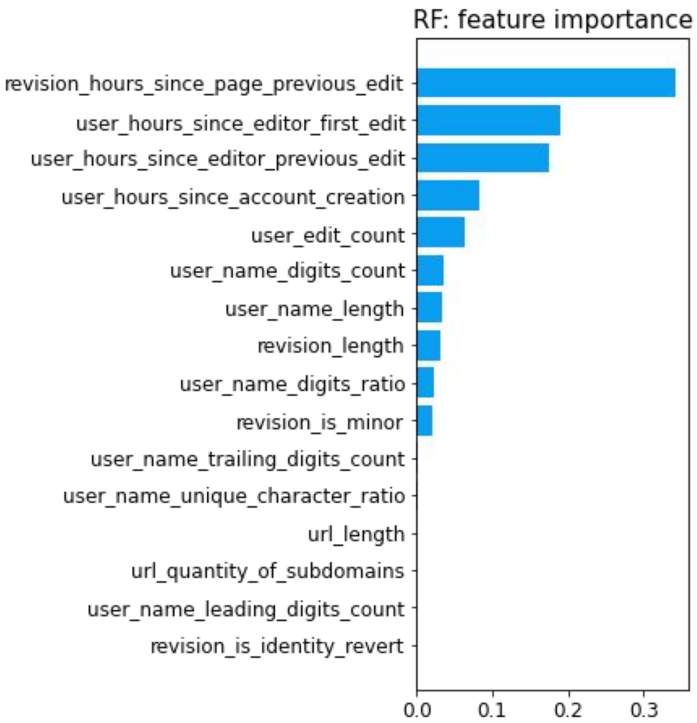


w.wiki/6649

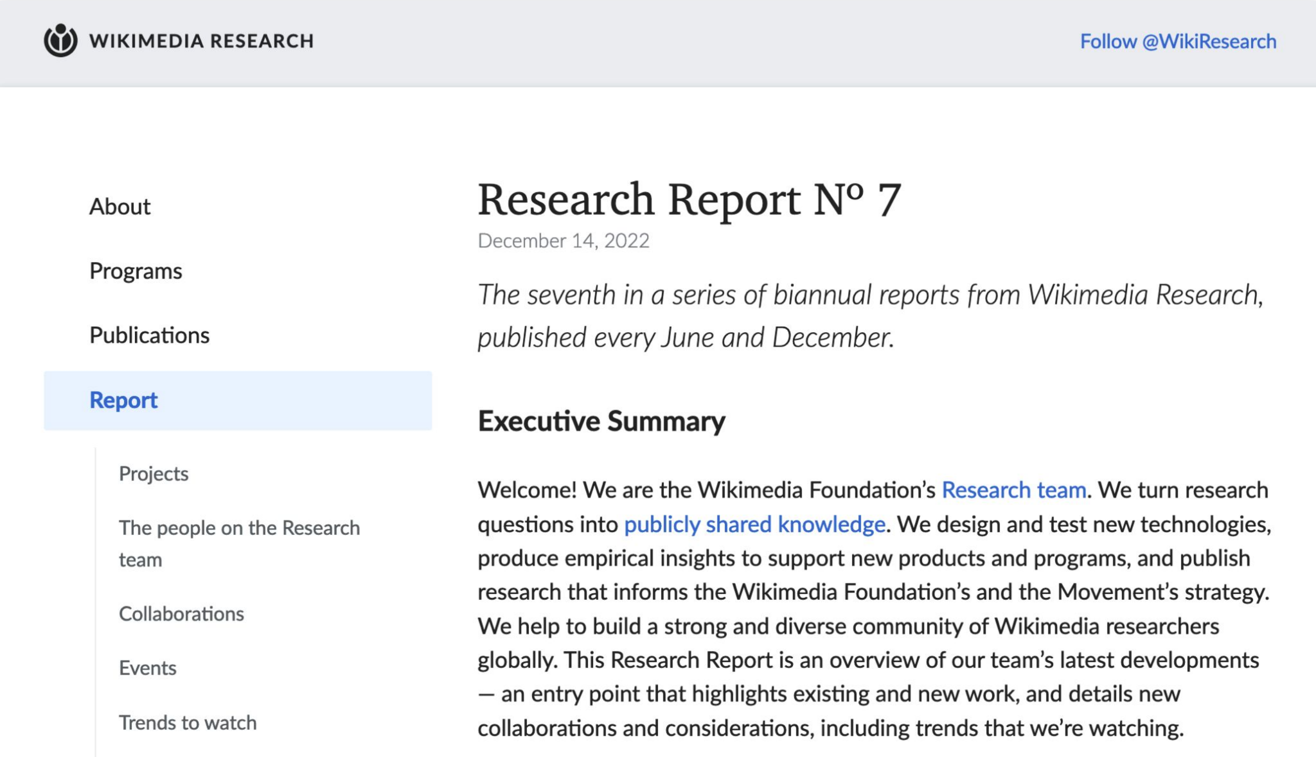
Examine needs for research

Spambot detection carried out by stewards





Read more on our Research Report!



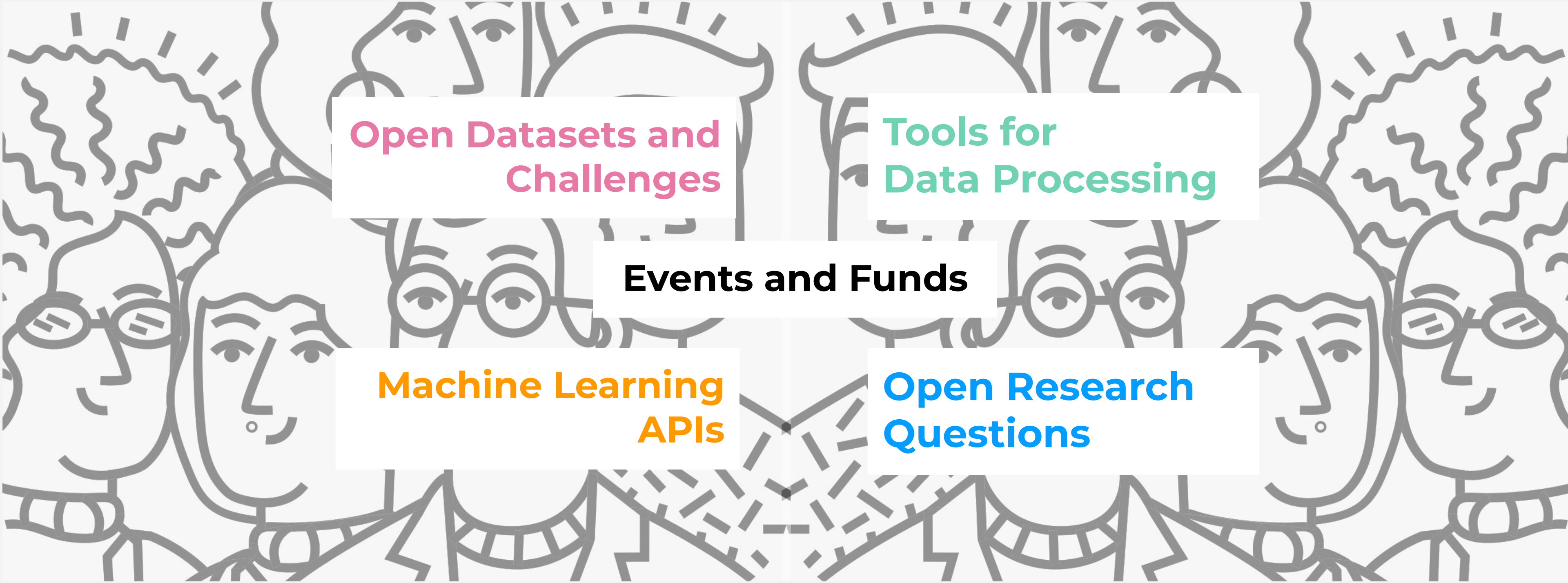
Research Report Nº 7

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.



Open Datasets and Challenges

Tools for Data Processing

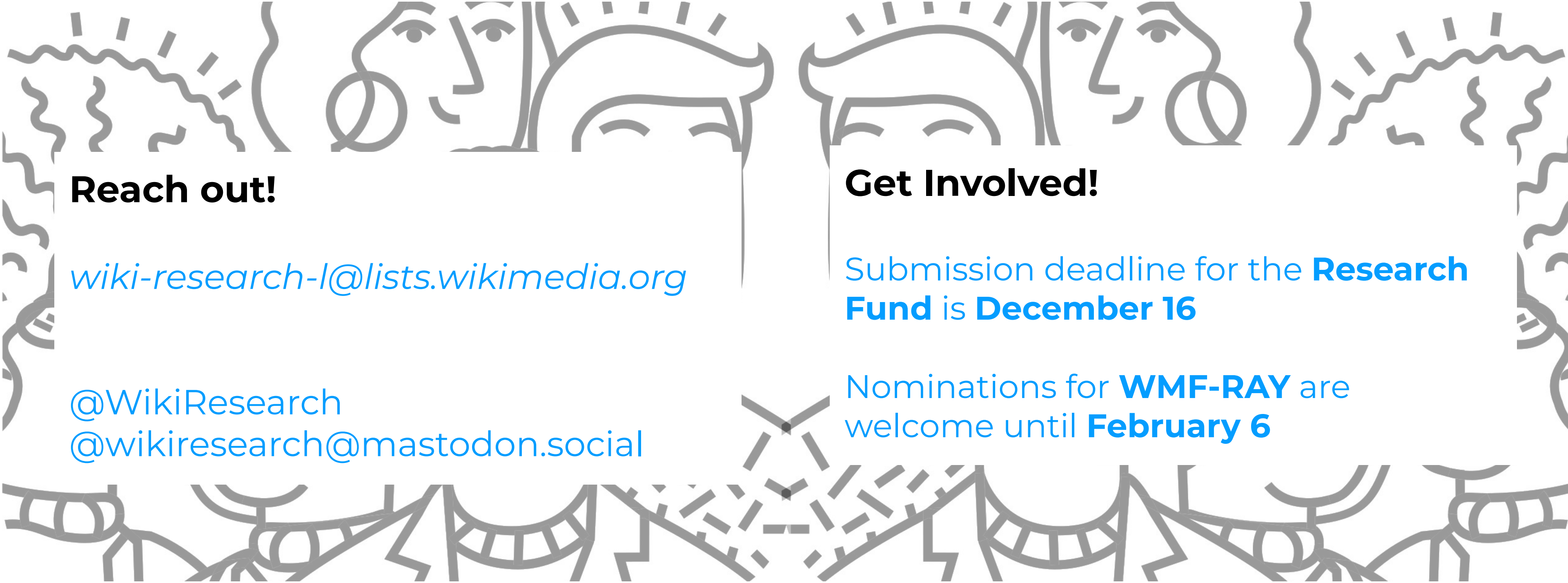
Events and Funds

Machine Learning APIs

Open Research Questions

DARIO TARABORELLI / CC0

Thank you!



Reach out!

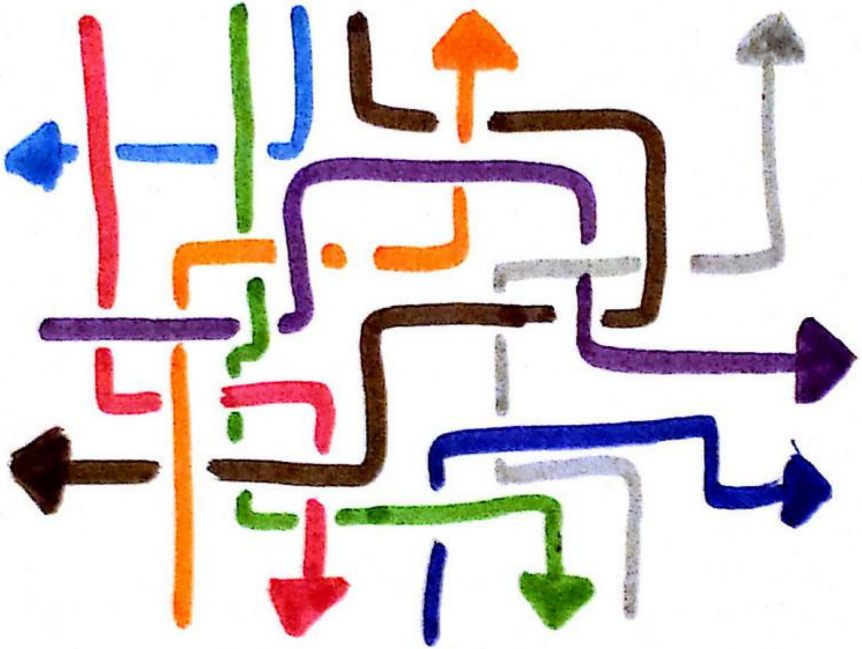
wiki-research-l@lists.wikimedia.org

@WikiResearch
@wikiresearch@mastodon.social

Get Involved!

Submission deadline for the **Research Fund** is **December 16**

Nominations for **WMF-RAY** are welcome until **February 6**



What excites you about your 2023?

Closing Slide

The Wikimedia Foundation’s Research Team