TECHNOLOGY DEPT

Community Support Programs

January 2018 quarterly check-in
for work done in Q2 FY2017/18

Program Structure

Sustaining	TP1 Availability, Performance & Maintenance	TP3 Addressing Technical Debt TP8 Multi-datacenter Support
Foundational	TP2 Mediawiki Refresh TP6 Streamlined Service Delivery PP1 Discoverability	X-SPDM Security, Privacy, & Data mgmt X-SDC Structured Data on Commons
Community Support	TP5 Scoring Platform (ORES) TP9 Growing Wikipedia Across Languages [...] TP7 Smart Tools for Better Data	TP11 Citations/Verifiability TP12 Growing Contributor Diversity X-CH Community Health/Anti-harassment
Tech Community Support	TP4 Technical Community Building	TP10 Public Cloud Services & Support

Program Priorities

	If we don't do this ...	Actual FTE (approximate)
Sustaining	The sites go down.	30
Foundational	Performance and data quality decays.	10
Community Support	Become technologically obsolete.	10
Tech Community Support	Lose bots and code contributions.	4

How we prioritize

Fundamentals

What is our part in fulfilling the mission?

Service

What are other people asking from us?

Improvement

What could we do to improve our offering?

Maintenance

What will sustain and improve our delivery?

Agenda:

Technology

- Program 4: Technical community building
- Program 5: Scoring Platform (ORES)
- Program 7: Smart tools for better data
- Program 9: Growing Wikipedia across languages
- Program 10: Public cloud services and support
- Program 11: Improving citations across Wikimedia projects

Programs covered in other presentations:

- Community Health (Research)
- Structured Data on Commons (Programs)

Technology Program 4

Technical community building

Tech Community Support	TP4 Technical Community Building

Wikimedia's software products and platforms have a diverse collection of technical and non-technical communities that have not always been well recognized for their contributions and supported in their work.

We will address this shortcoming by providing better documentation, facilitating community building, and establishing better pathways for communication between these communities and the Foundation.

Wikimedia Hackathon badge stickers By Wikimedia Österreich

Outcome 1 / Objective 2: Create tutorial content for common issues

1. Overview
2. Getting started
3. Steps to completion
4. Step-by-step guide
5. Step 1: Create a new tool account
6. Step 2: Create a basic Flask WSGI webservice
7. Step 3: Add a configuration file
8. Step 4: Add support for OAuth authentication
9. Additional troubleshooting
10. Next Steps
11. See also

WIKIMEDIA FOUNDATION

Outcome 2:

The adoption of Wikimedia technology can be reliably measured.

Photo by Ron Smith on Unsplash

Outcome 3/Objective 1:

Establish ongoing channels of communication with third-party developers

Participate in:

- MediaWiki Stakeholders Group
  - user group recognized by Affiliations Committee consisting of developers, system administrators, and users who cooperate to improve the software and advocate for MediaWiki users outside Wikimedia projects
- Third-party developer conferences:
  - Semantic MediaWiki Conference (SMWCon)
  - Enterprise MediaWiki Conference (EMWCon)
- Enterprise MediaWiki Slack team (NASA)
  - Q3: encourage use of FOSS tool to bridge communication with Wikimedia communities
- Monthly US Federal Government MediaWiki telecon

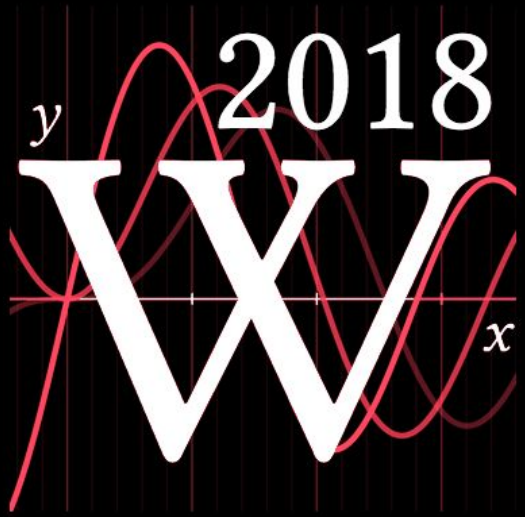
Encourage third parties to participate in Wikimedia events:

- Developer Summit (NASA, Hallo Welt!/BlueSpice, NicheWork)

Outcome 3/Objective 2:

Clarify the Foundation's short- and long-term commitments to third-party users

- Develop strategy for MediaWiki for third party usage *(in progress)*
  - Collecting information from current and former third-party MediaWiki users
- Q3: Create and publish a multi-tiered, third-party support level system for Wikimedia extensions frequently used by third parties



Wiki Workshop 2018

The Web Conference 2018
Lyon • 24 April 2018

THE WEB CONFERENCE

wikiworkshop.org

2 days
4 keynotes
7 guests
8 topic sessions
55 invited attendees
= priceless
communication and information sharing with input for the movement strategy process

phabricator.wikimedia.org/T185012
mw:Wikimedia_Developer_Summit/2018
c:Category:Wikimedia_Developer_Summit_2018

Tech Managers

Face to Face

January 24, 2018

Discussion highlights:

- annual planning
- free software ideas
- tech culture

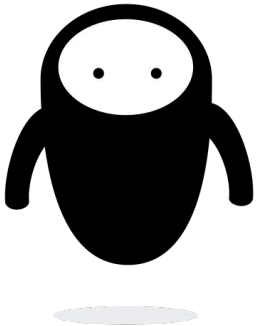
Technology Program 5

Scoring Platform (ORES)

Program Structure

Community Support	TP5 Scoring Platform (ORES) TP9 Growing Wikipedia Across Languages [...] TP7 Smart Tools for Better Data

We will help increase the efficiency of production activities on the wikis with machine prediction services and we will build accountability mechanisms to mitigate the effects of prediction errors and bias



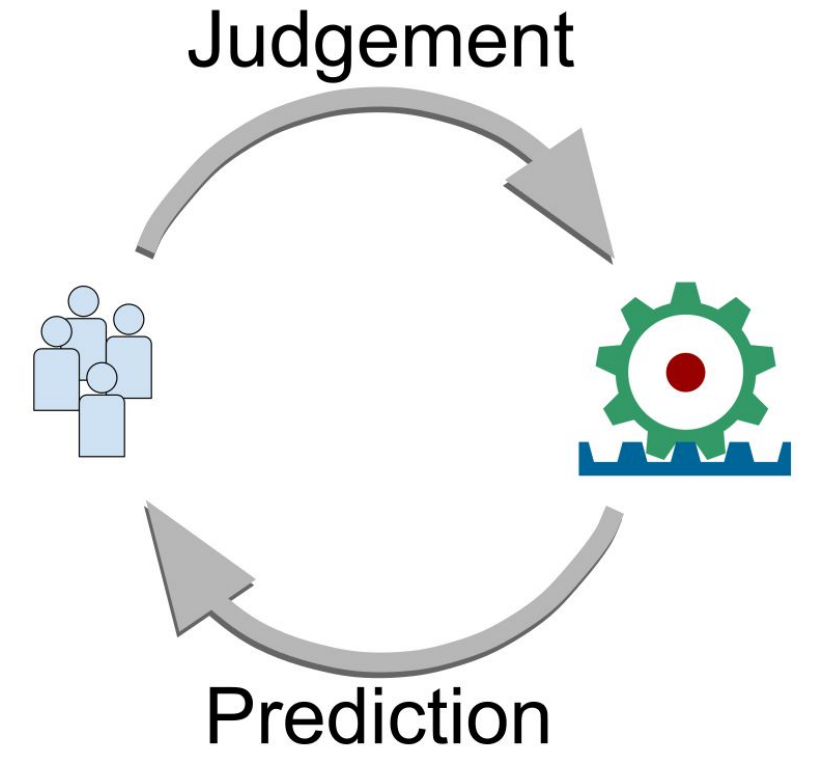
SCORING PLATFORM TEAM

Expanded counter-vandalism

WIKIPEDIA
Den fria encyklopedin

WIKIPEDIA
La enciclopedia libre

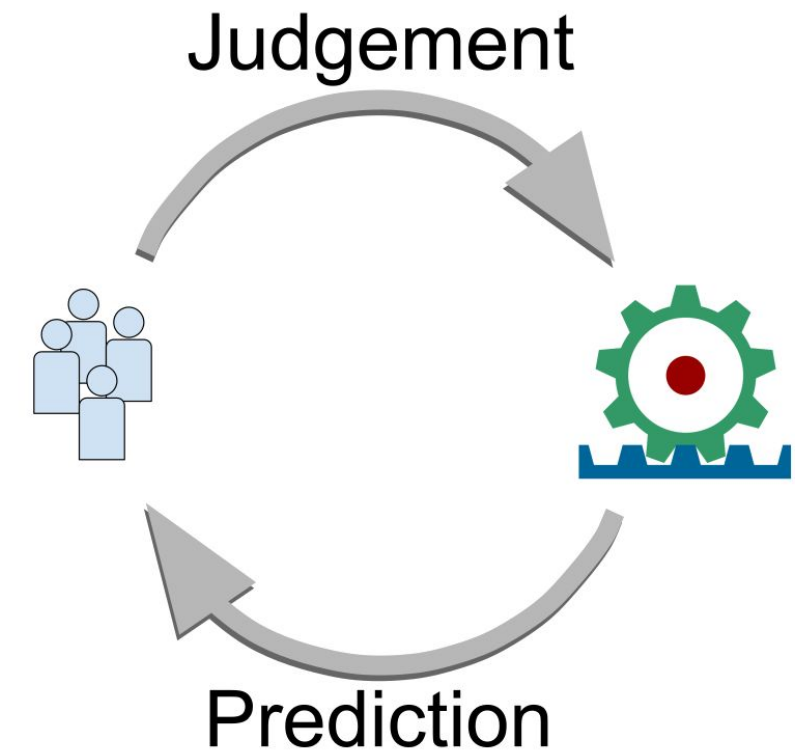
JADE: Judgement and Dialog Engine



- Auditing (Counter-bias)
- Coordination between patrollers
- Training new AIs

Read more at [[:mw:JADE]]

JADE: Judgement and Dialog Engine



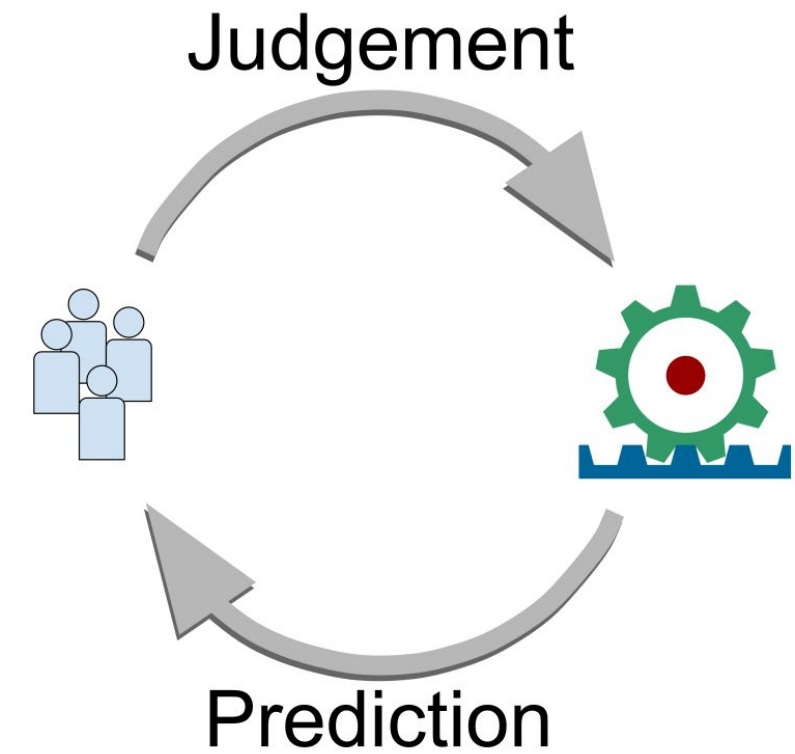
Free text comments and suppression

27 comments • 22 days ago

Summary by EpochFail

See task T183276 "Design curation/suppression integration with MediaWiki (for JADE)"

JADE: Judgement and Dialog Engine



Free text comments and suppression

27 comments • 22 days ago

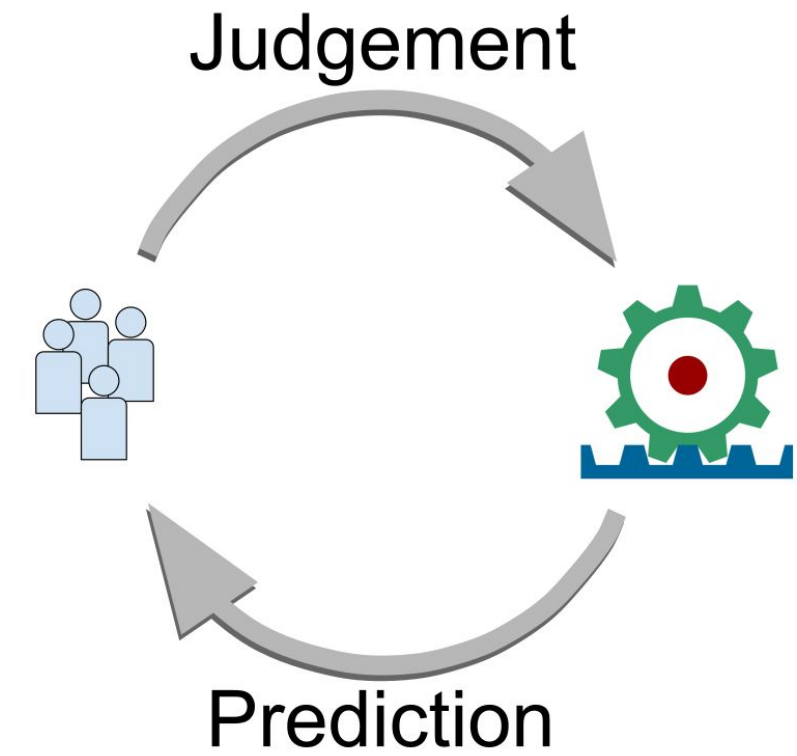
Summary by EpochFail

See task T183276 "Design curation/suppression integration with MediaWiki (for JADE)"

Judgment Bold-Revert-Discuss vs. optional consensus

6 comments • a month ago

JADE: Judgement and Dialog Engine



Free text comments and suppression
27 comments • 22 days ago

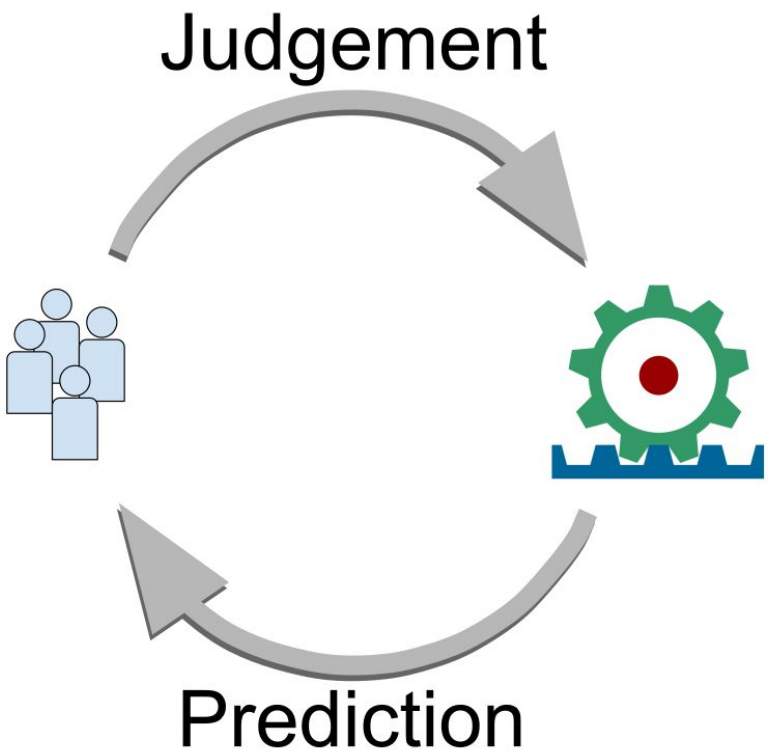
Summary by EpochFail

See task T183276 "Design curation/suppression integration with MediaWiki (for JADE)"

Judgment Bold-Revert-Discuss vs. optional consensus
6 comments • a month ago

Judgments, Endorsements, and Preference
3 comments • 2 months ago

JADE: Judgement and Dialog Engine



Free text comments and suppression

27 comments • 22 days ago

Summary by EpochFail

See task T183276 "Design curation/suppression integration with MediaWiki (for JADE)"

Judgment Bold-Revert-Discuss vs. optional consensus

6 comments • a month ago

Judgments, Endorsements, and Preference

3 comments • 2 months ago

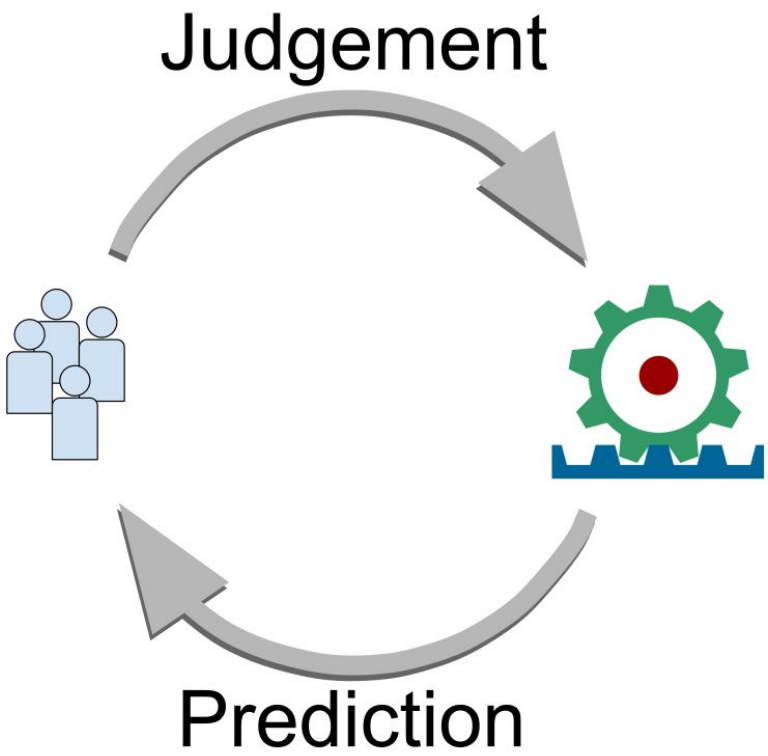
Thematic and quant analysis of judgements

3 comments • 2 months ago

Summary by Adamw

Judgments and comments should be available for analysis from Quarry, and joinable against wiki entities. Comments should be full-text searchable.

JADE: Judgement and Dialog Engine



Free text comments and suppression

27 comments • 22 days ago

Summary by EpochFail

See task T183276 "Design curation/suppression integration with MediaWiki (for JADE)"

Judgment Bold-Revert-Discuss vs. optional consensus

6 comments • a month ago

Judgments, Endorsements, and Preference

3 comments • 2 months ago

Thematic and quant analysis of judgements

3 comments • 2 months ago

Summary by Adamw

Judgments and comments should be available for analysis from Quarry, and joinable against wiki entities. Comments should be full-text searchable.

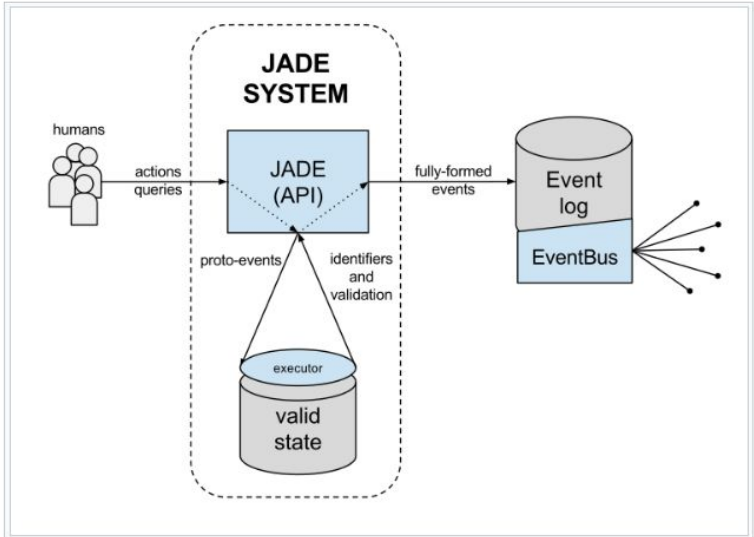
Should we integrate JADE with Structured Discussions?

11 comments • 2 months ago

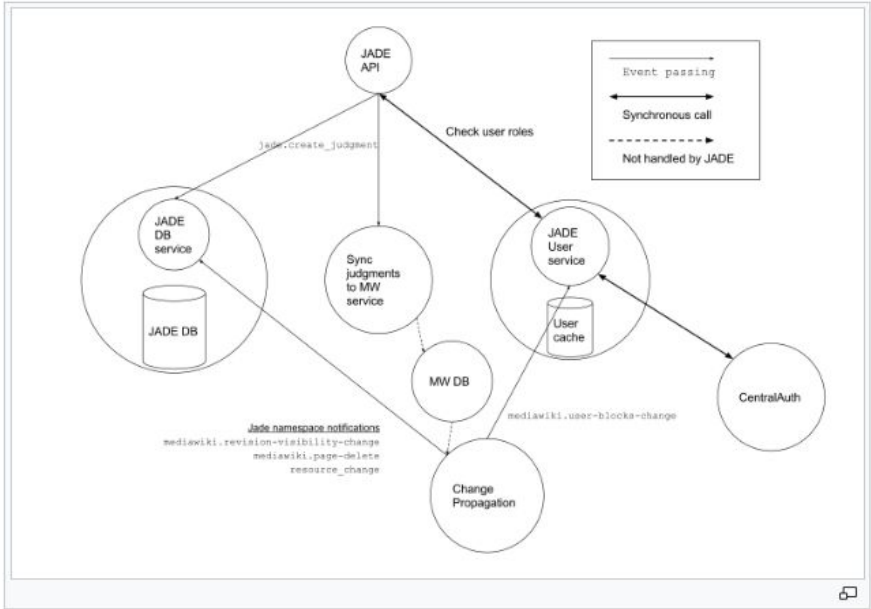
Summary by Adamw

Judgments and discussions are rooted at wiki entities. We need more discussion before committing to using Structured Discussions, it's not the perfect fit. Having ORES and JADE entities available when reading or editing these threads would be great.

[[mw:Talk:JADE/Implementations]]



JADE state validator. The process of generating events is expanded to include a flow of proto-events and responses from a database that then results in emitting fully-formed events to the event log.



Free text comments and suppression

27 comments • 22 days ago

Summary by EpochFail

See task T183276 "Design curation/suppression integration with MediaWiki (for JADE)"

EpochFail

There's one thing that's very clear from reading through ORES' mistake reports online: People want to include freetext comments with their judgements.

However, there's a problem where any time you open up a new freetext field on the internet, someone's going to use it to Dox someone else or otherwise cause harm. So we're going to need some mechanism for curation and suppression. There are two options we're kicking around.

First class suppression support

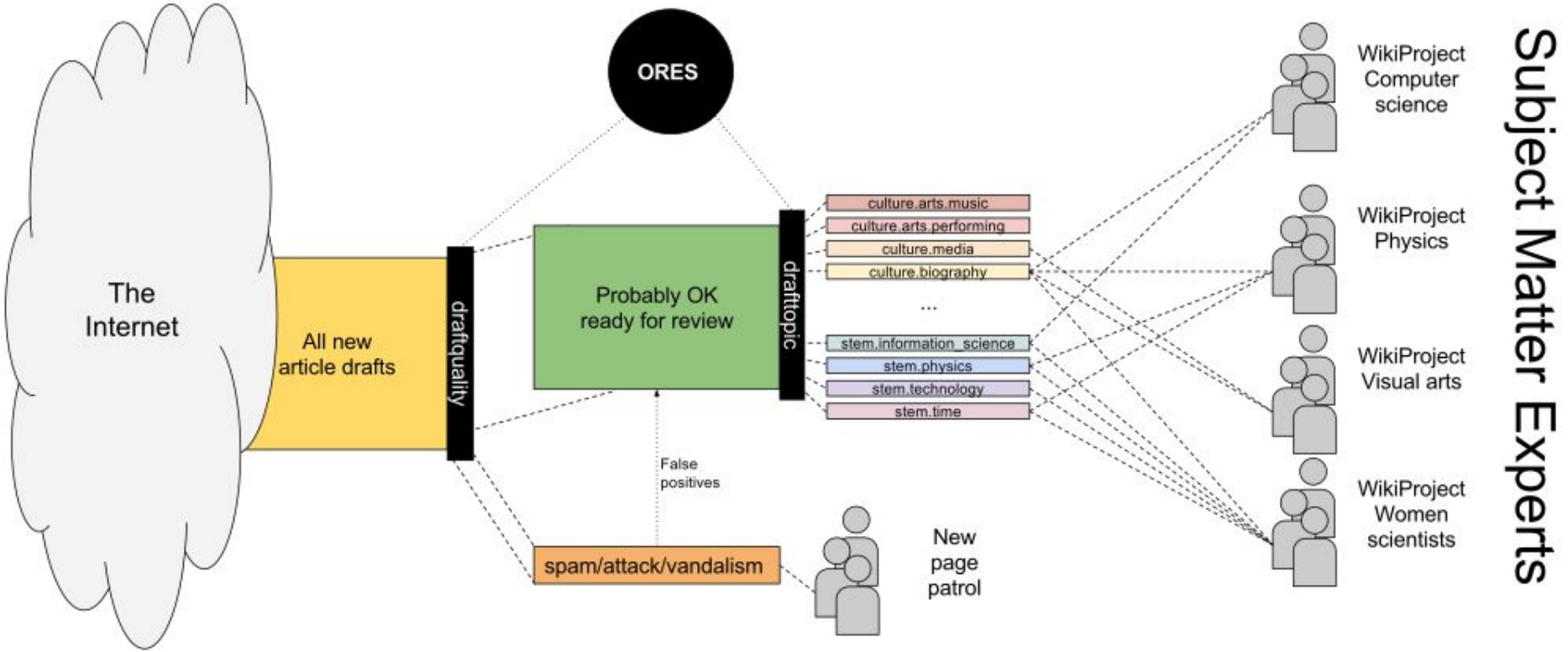
Text fields would be stored within JADE and a full suite of suppression tools would be made available. Integrations with MediaWiki would allow patrollers to see JADE comments appear in the RecentChanges feed on their local wiki.

Upside

Text is part of the system and any analysis users will want to do.

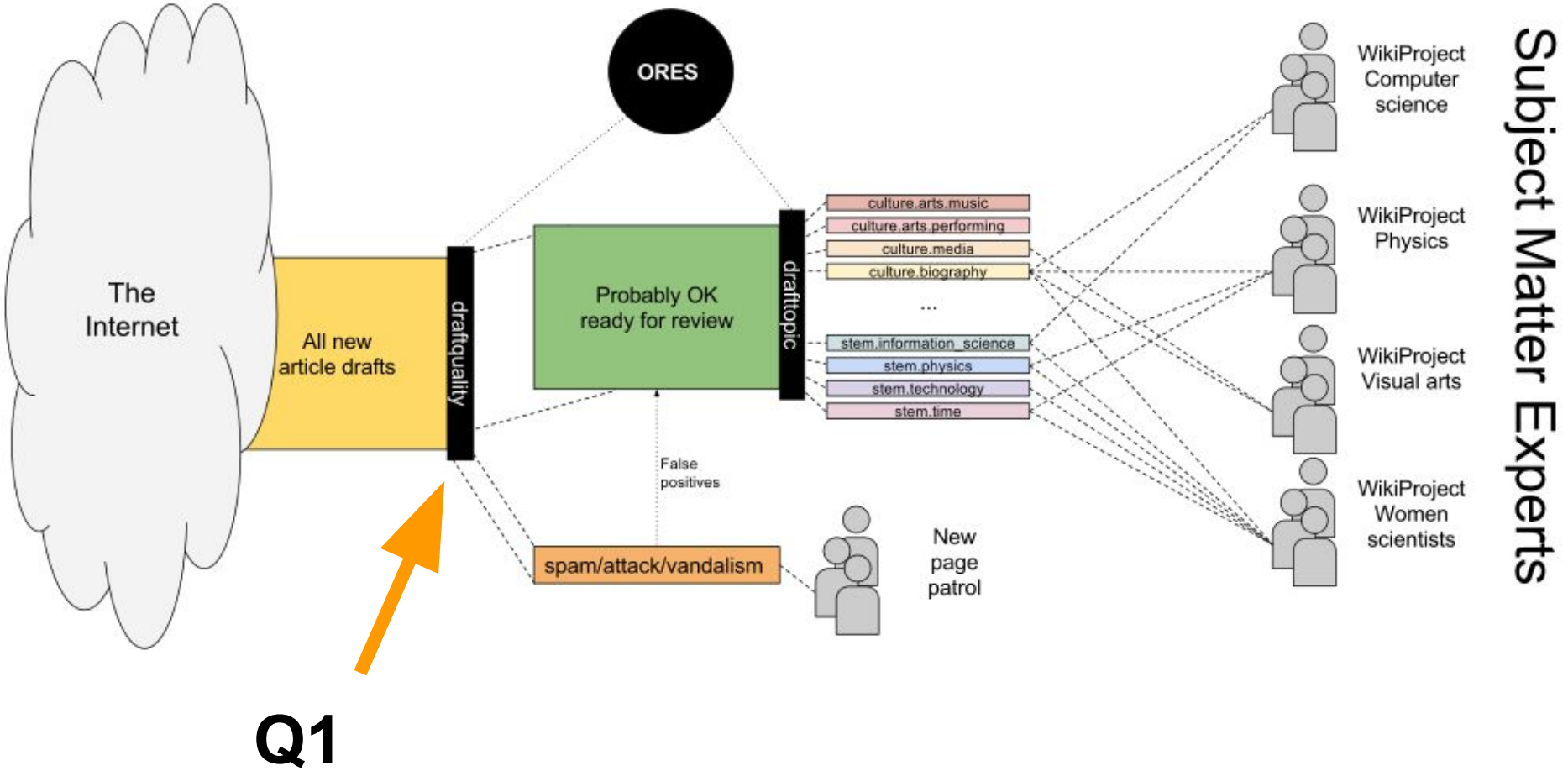
Downside

Way more work to implement and easy to get wrong.

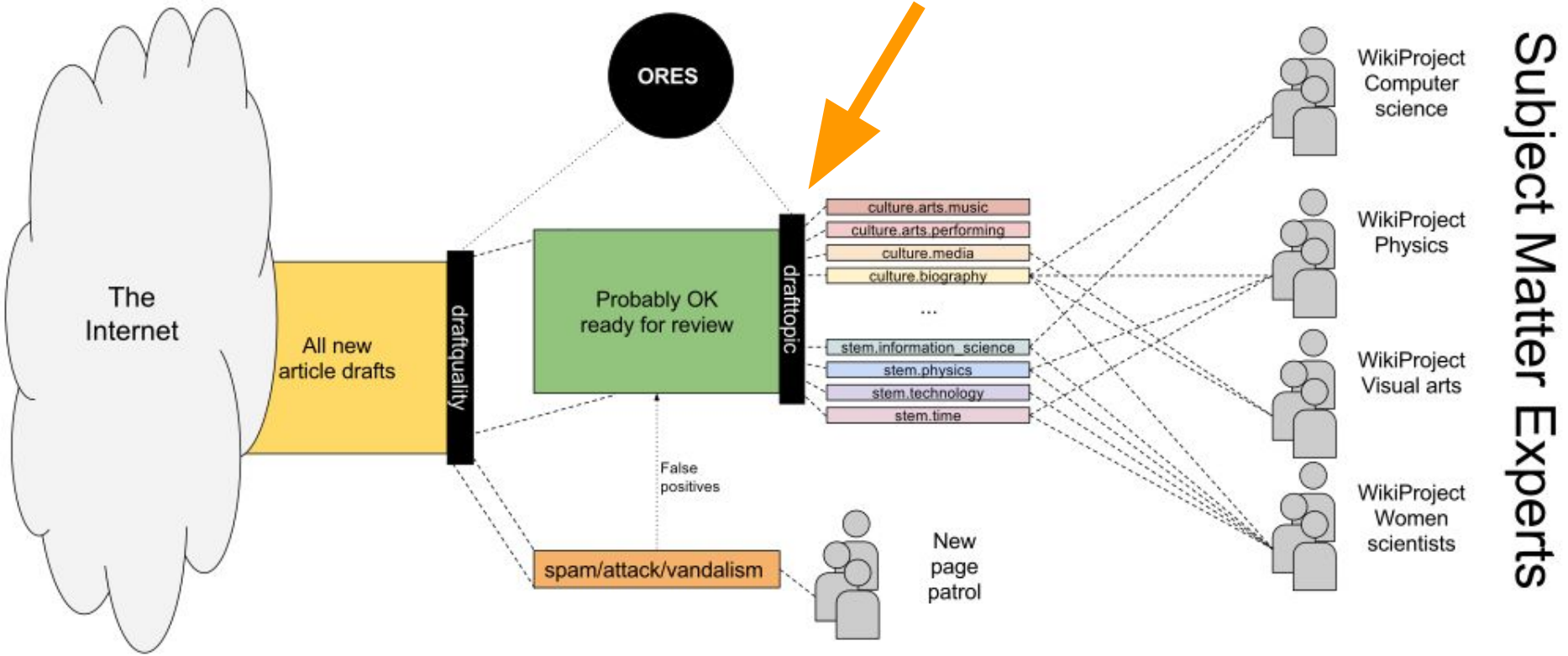

The Internet
All new article drafts
draftquality
ORES
Probably OK ready for review
drafttopic
culture.arts.music
culture.arts.performing
culture.media
culture.biography
...
stem.information_science
stem.physics
stem.technology
stem.time
False positives
spam/attack/vandalism
New page patrol
WikiProject Computer science
WikiProject Physics
WikiProject Visual arts
WikiProject Women scientists
Subject Matter Experts

[[:mw:Wikimedia_Research/Showcase]]

February 21st, 2018


ORES
The Internet
All new article drafts
draftquality
Probably OK ready for review
drafttopic
culture.arts.music
culture.arts.performing
culture.media
culture.biography
...
stem.information_science
stem.physics
stem.technology
stem.time
False positives
spam/attack/vandalism
New page patrol
WikiProject Computer science
WikiProject Physics
WikiProject Visual arts
WikiProject Women scientists
Subject Matter Experts
Q1

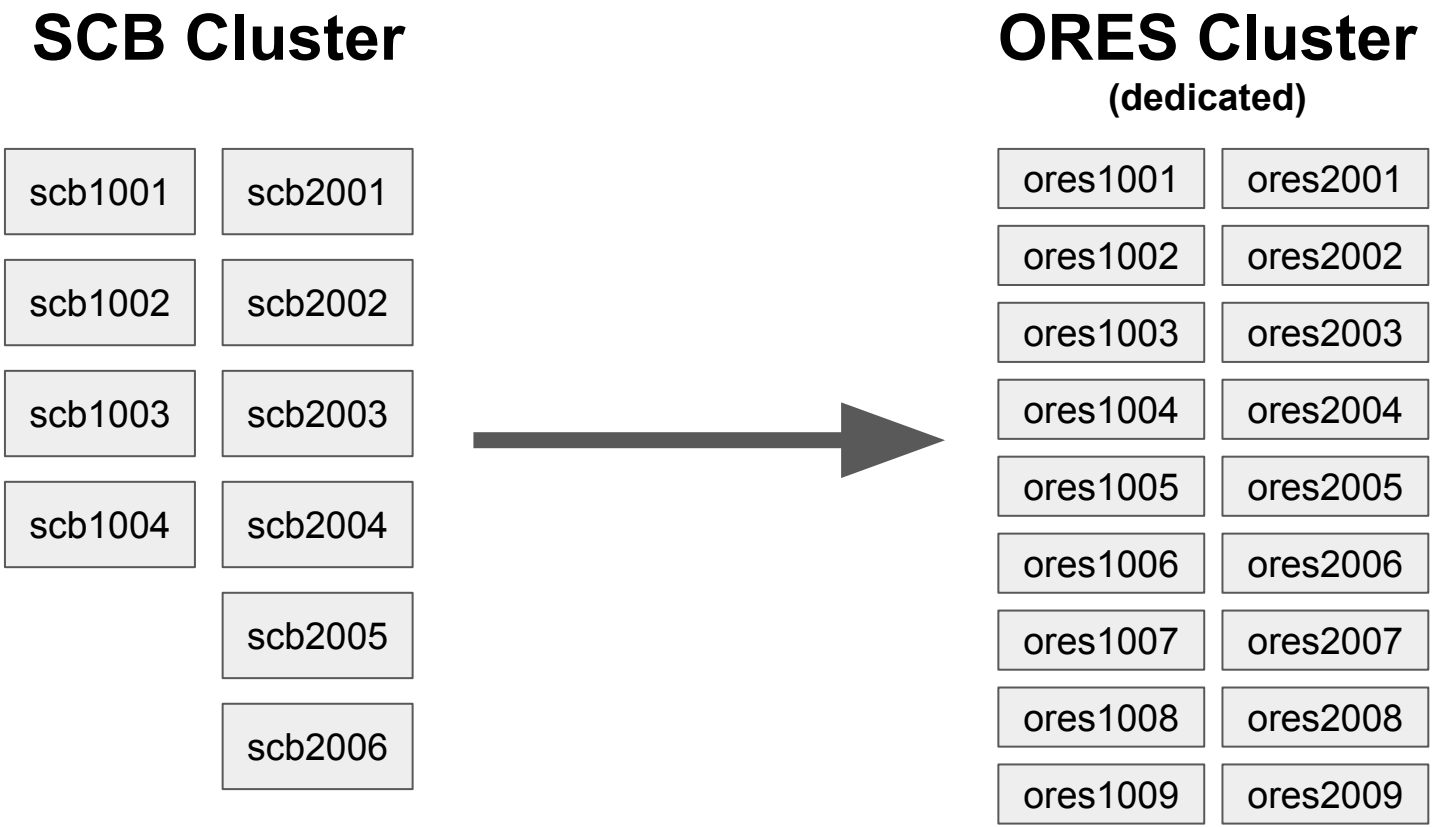
Q2 - Q3


ORES
The Internet
All new article drafts
draftquality
Probably OK ready for review
drafttopic
culture.arts.music
culture.arts.performing
culture.media
culture.biography
...
stem.information_science
stem.physics
stem.technology
stem.time
False positives
spam/attack/vandalism
New page patrol
WikiProject Computer science
WikiProject Physics
WikiProject Visual arts
WikiProject Women scientists
Subject Matter Experts

SCB Cluster

scb1001	scb2001
scb1002	scb2002
scb1003	scb2003
scb1004	scb2004
	scb2005
	scb2006

- ORES
- changeprop
- citoid
- cpjobqueue
- cxserver
- eventstreams
- graphoid
- mathoid
- mobileapps
- pdfrender (scb)


SCB Cluster
scb1001
scb2001
scb1002
scb2002
scb1003
scb2003
scb1004
scb2004
scb2005
scb2006
ORES Cluster
(dedicated)
ores1001
ores2001
ores1002
ores2002
ores1003
ores2003
ores1004
ores2004
ores1005
ores2005
ores1006
ores2006
ores1007
ores2007
ores1008
ores2008
ores1009
ores2009

Outages

- November 20th
- November 28th

Actionables [edit | edit source]

- ✓ **Done** bug T181191 - Make MediaWiki pages robust to ORES or Ext:ORES failures.
- ✓ **Done** bug T181183 - Deployment documentation and protocol to cover what awight missed here.
- (pending review) bug T181071 - Cache virtualenv for faster rollback.
- ✓ **Done** bug T181067 - Parallelize ORES deployment.
- ✓ **Done** bug T181187 - Always make ORES beta cluster config the same as production.

Outreach

- New ORES FAQ (Sarah Rodlund & Technical contributors)

University visits

- Idaho State
- University of Michigan
- Stanford
- University of Washington
- Roles in open collaboration Workshop
- AI Bias and Mitigation workshop

Conferences & Keynotes

- ACM GROUP (Social Science & Computer Science)
- International Wiki Science Conference

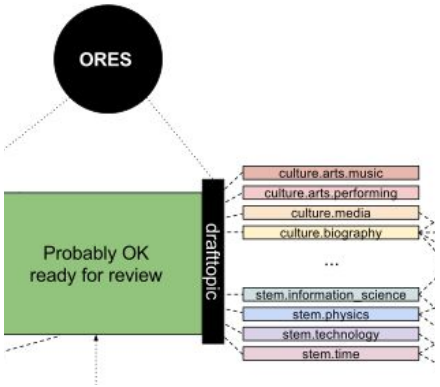
FY18: 3rd Quarter

- ORES extension refactoring -- no more avoidable outages!
- Expand counter-vandalism support

Simple English WIKIPEDIA

WIKIPÉDIA
A szabad enciklopédia

- Deploy the “draft topic” model in ORES
- JADE discussions → Blog
- JADE MVP attempt #2



Thanks!

Team [edit | edit source]

Aaron Halfaker
Principal Research Scientist
Team Lead

Amir Sarabadani
Software Engineer (WMDE)

Adam Wight
Software Engineer (WMF)

Sumit
Contractor

1 + 0.5 + 0.8 + 0.25

= 2.55

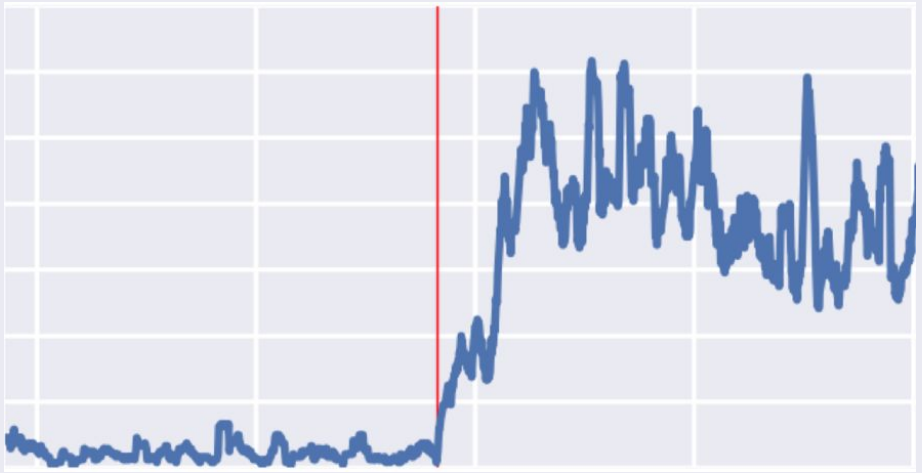
Technology Program 7

Smart tools for better data

Community Support

TP5 Scoring Platform (ORES)
TP9 Growing Wikipedia Across Languages [...]
TP7 Smart Tools for Better Data

We will maintain and increase public access to past, present and real time data for Wikimedia projects. We will provide the infrastructure to measure the impact and reach of projects and features for editors, communities and WMF.



Wikistats 2.0 (Alpha)

Yes, a website: http://stats.wikimedia.org/v2

Wikistats 2.0 (Alpha)

Yes, a website: http://stats.wikimedia.org/v2

But also...

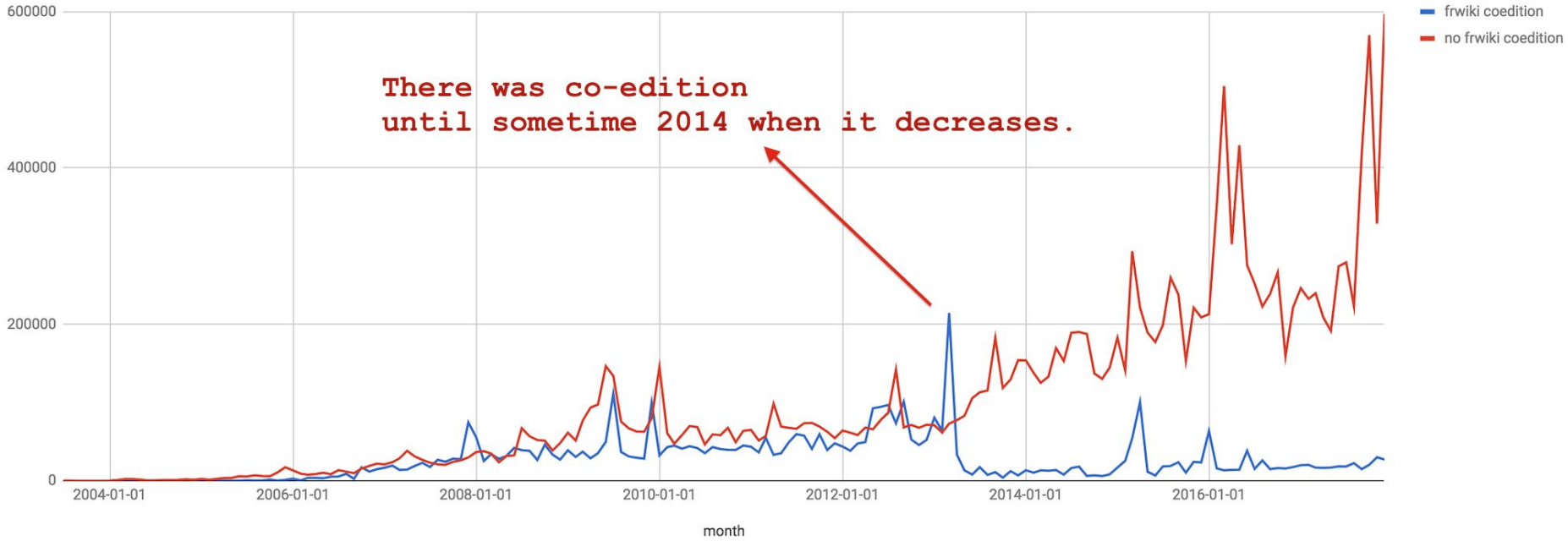
The BEST dataset we ever had to answer questions about CONTENT and CONTRIBUTORS. Updated Monthly.


The BEST dataset we ever had to answer questions about CONTENT and CONTRIBUTORS. Updated Monthly.
a.k.a. Data Lake

The BEST dataset we ever had to answer questions about CONTENT and CONTRIBUTORS.

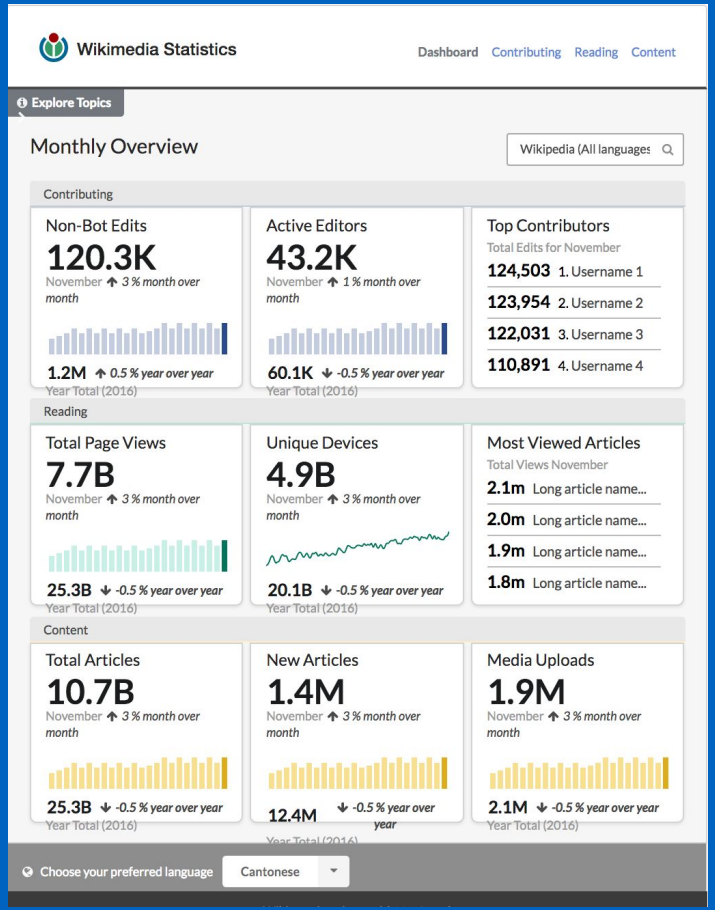
Example: How many edits are done per month in Arabic Wikipedia by editors that also edit French Wikipedia since 2005?

How many edits are done per month in Arabic Wikipedia by editors that also edit French Wikipedia since 2005?



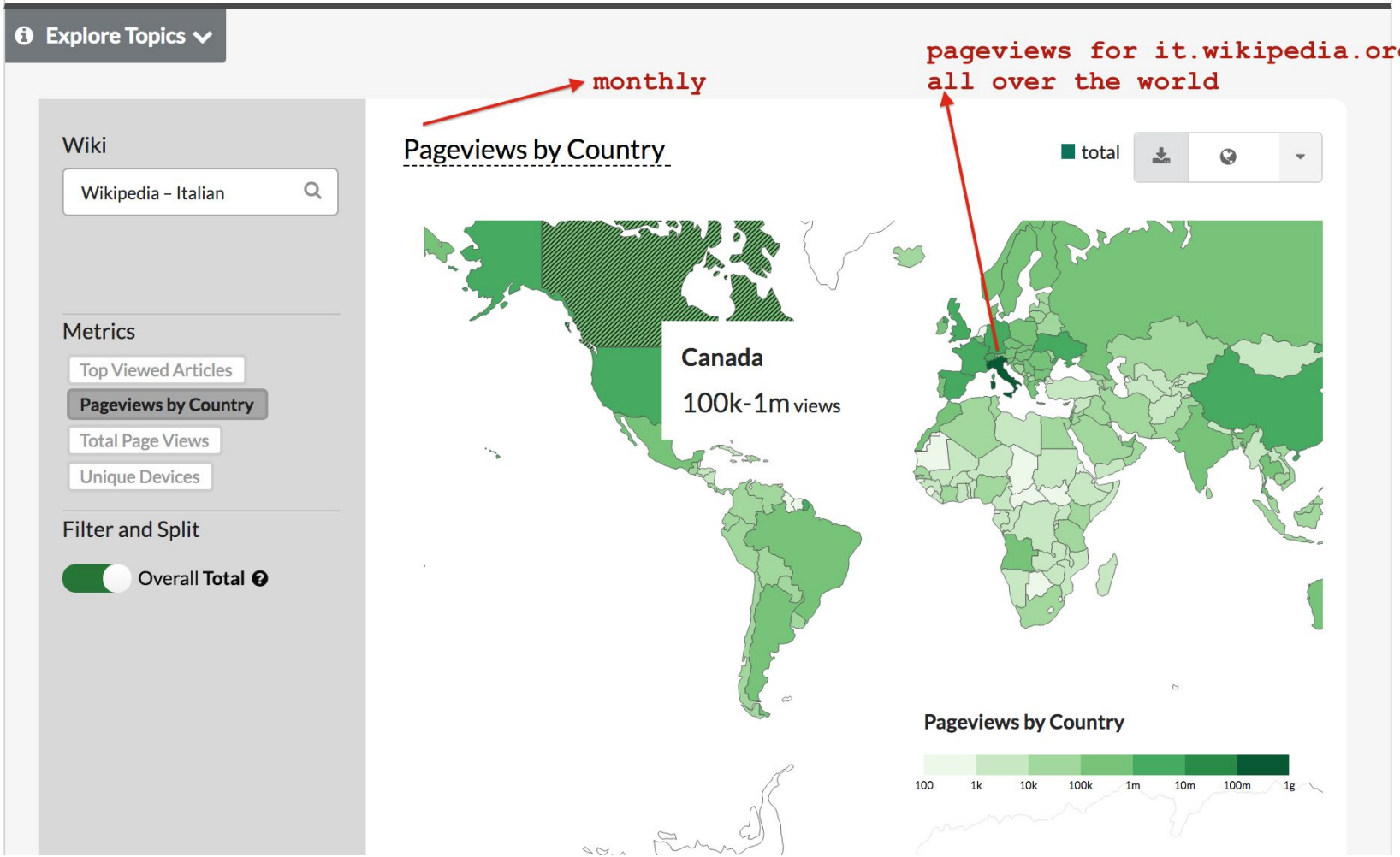
Wikistats exists to motivate our editor community.

In Wikistats 2.0 we are not only updating the website interface but we are also providing new access to all our edit data in an analytics-friendly form. This greatly improves (and fundamentally changes) the way, time and resources it takes to calculate edit metrics, for WMF and community.



WIKIMEDIA FOUNDATION

Sneak Preview..


Explore Topics
monthly
pageviews for it.wikipedia.org
all over the world
Wiki
Wikipedia – Italian
Metrics
Top Viewed Articles
Pageviews by Country
Total Page Views
Unique Devices
Filter and Split
Overall Total
Pageviews by Country
total
Canada
100k-1m views
Pageviews by Country
100
1k
10k
100k
1m
10m
100m
1g

Superset. Better visual access to data for WMF PMs

http://superset.wikimedia.org

You too can build DASHBOARDS!

Pageviews Overview ☆

Please notice it
has dynamic components

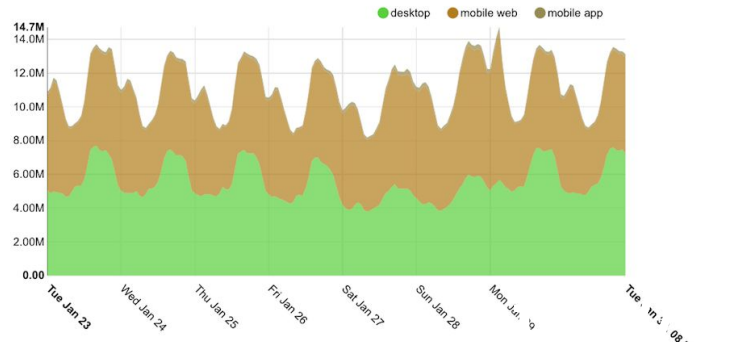
pageviews-project-filter

project

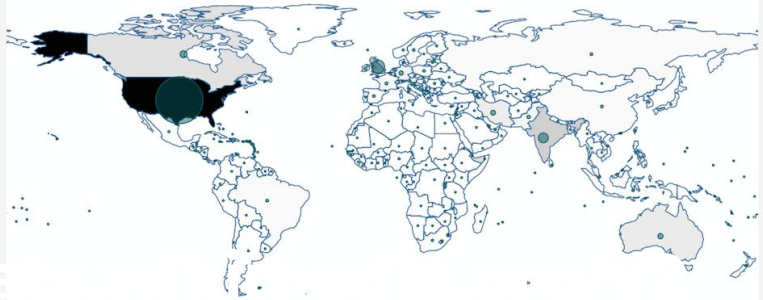
× en.wikipedia

pageviews-yesterday

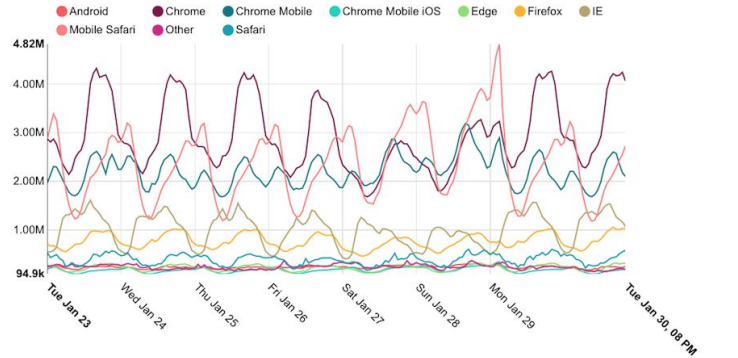
pageviews-access-hourly-currentweek



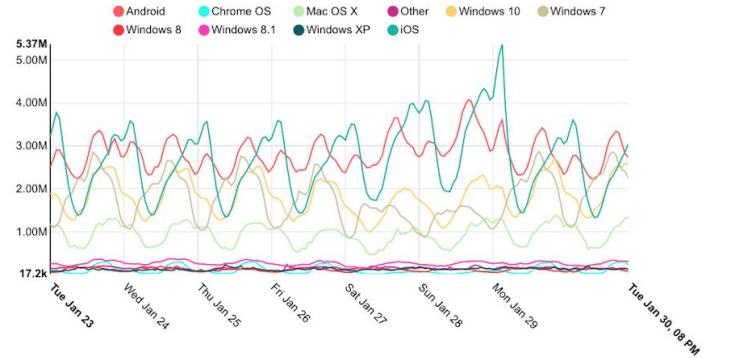
pageviews-countries-yesterday



pageviews-browserfamily-hourly-currentweek



pageviews-osfamily-hourly-currentweek



Pageviews Overview ☆

Please notice it
has dynamic components

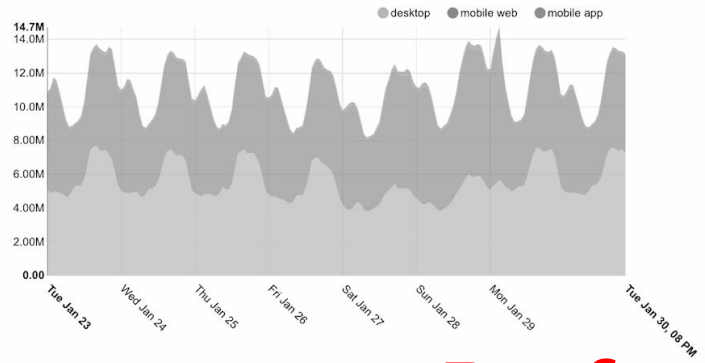
pageviews-project-filter

project

× en.wikipedia

pageviews-yesterday

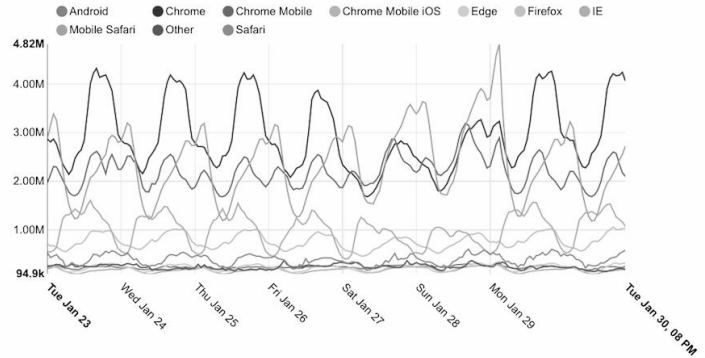
pageviews-access-hourly-currentweek



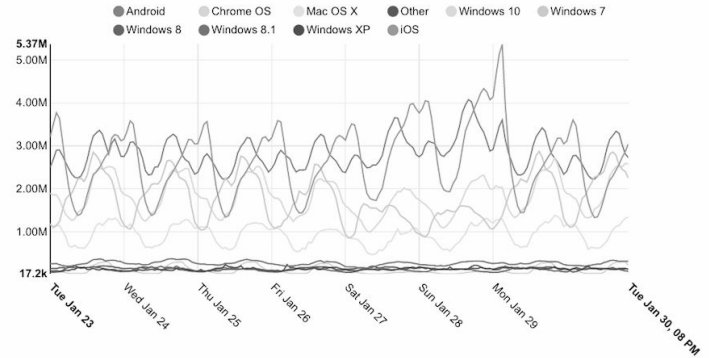
pageviews-countries-yesterday

Performance?

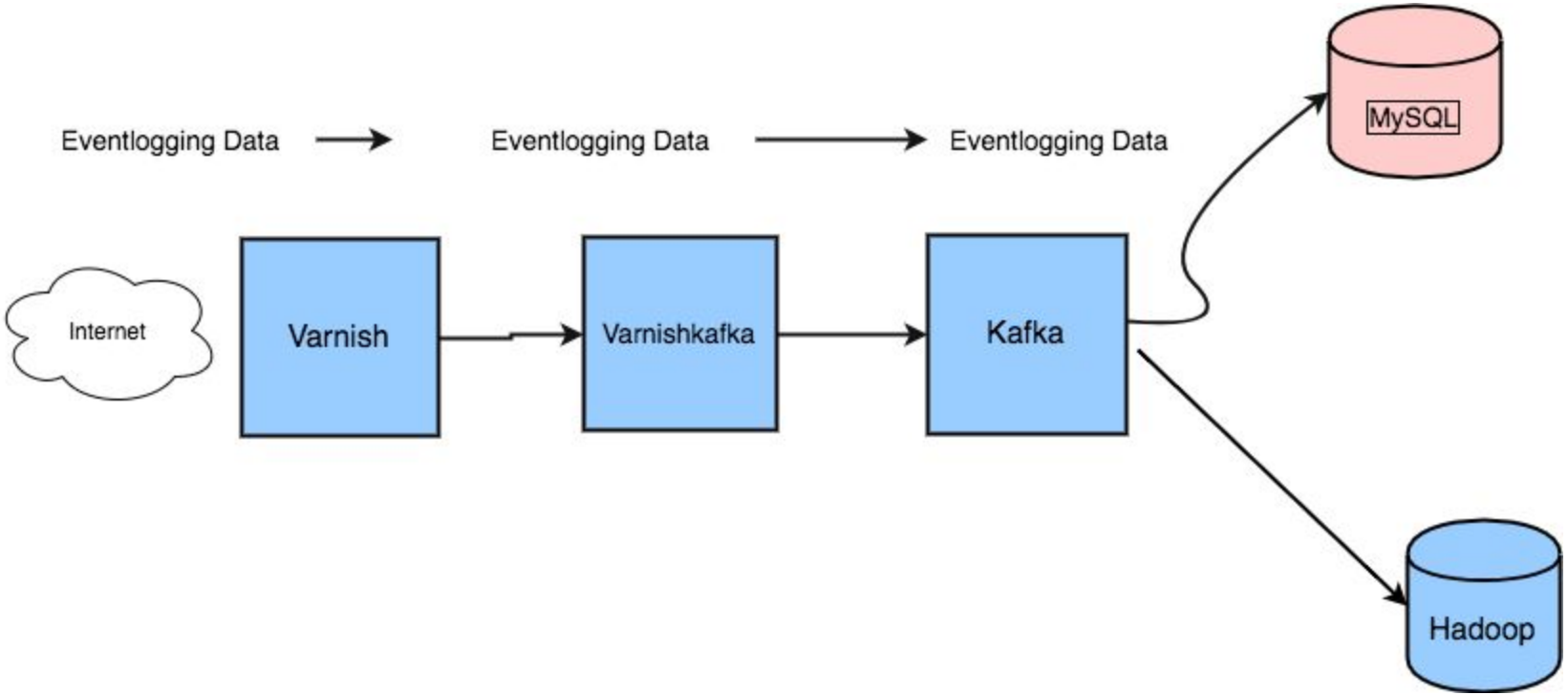
pageviews-browserfamily-hourly-currentweek



pageviews-osfamily-hourly-currentweek



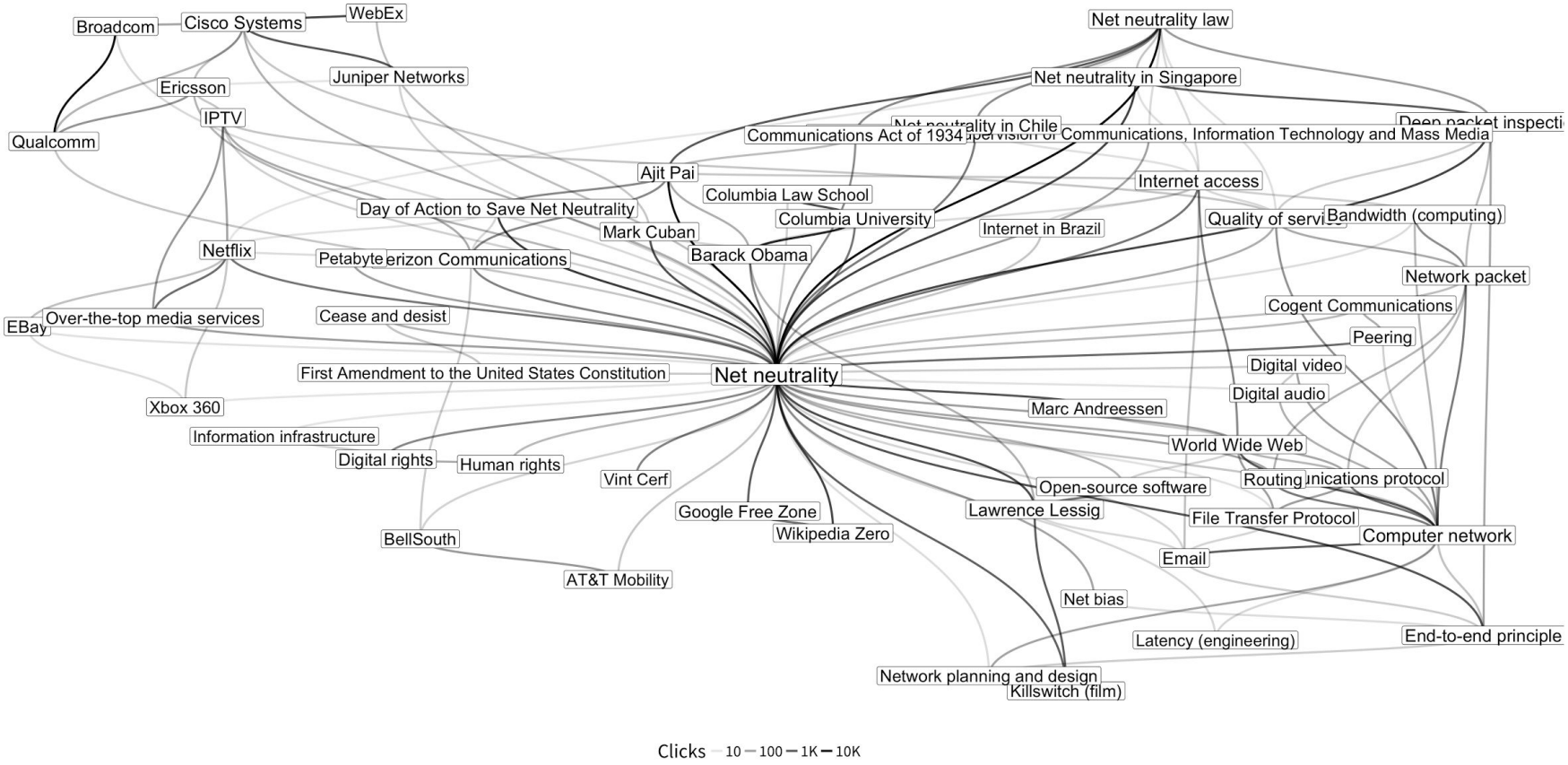
Scalable (and friendly) backend for Eventlogging. Hadoop. Druid.


Eventlogging Data
Eventlogging Data
Eventlogging Data
MySQL
Internet
Varnish
Varnishkafka
Kafka
Hadoop

New Dataset: Clickstream.

New Dataset: Clickstream.

How our users (collectively, not individually) traverse wikipedia.


Broadcom
Cisco Systems
WebEx
Juniper Networks
Ericsson
IPTV
Qualcomm
Net neutrality law
Net neutrality in Singapore
Net neutrality in Chile
Communications Act of 1934
Supervision of Communications, Information Technology and Mass Media
Ajit Pai
Columbia Law School
Columbia University
Day of Action to Save Net Neutrality
Mark Cuban
Barack Obama
Internet access
Internet in Brazil
Bandwidth (computing)
Netflix
Petabyte
Verizon Communications
Network packet
Cogent Communications
EBay
Over-the-top media services
Cease and desist
Peering
First Amendment to the United States Constitution
Net neutrality
Digital video
Digital audio
Xbox 360
Marc Andreessen
Information infrastructure
World Wide Web
Digital rights
Human rights
Vint Cerf
Routing
Open-source software
Google Free Zone
Lawrence Lessig
File Transfer Protocol
Wikipedia Zero
Computer network
BellSouth
Email
AT&T Mobility
Net bias
Latency (engineering)
End-to-end principle
Network planning and design
Killswitch (film)
Clicks
10
100
1K
10K

Technology Program 9

Growing Wikipedia across languages

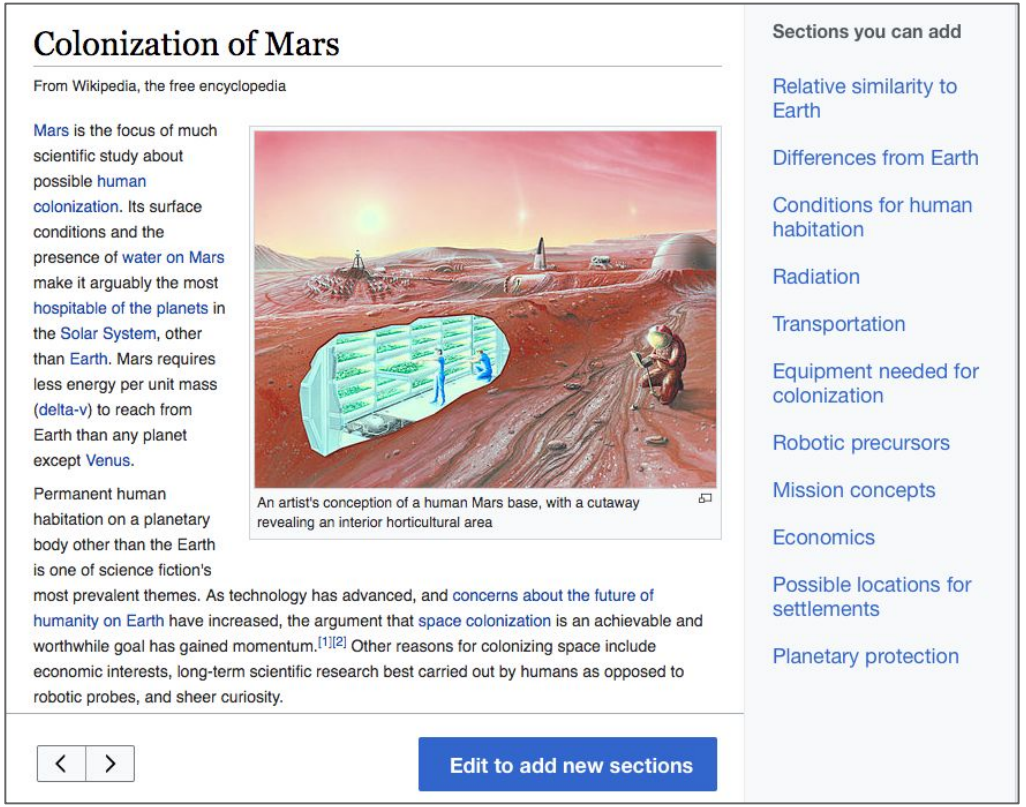
Program Structure

Community Support

TP5 Scoring Platform (ORES)
TP9 Growing Wikipedia Across Languages [...]
TP7 Smart Tools for Better Data

What is the problem?

There is no canonical structure for Wikipedia articles, across topics and languages.



Why is this a problem?

Onboarding newcomers at scale is hard.

14,000 new accounts get created every month.
Many need help to learn how to contribute.

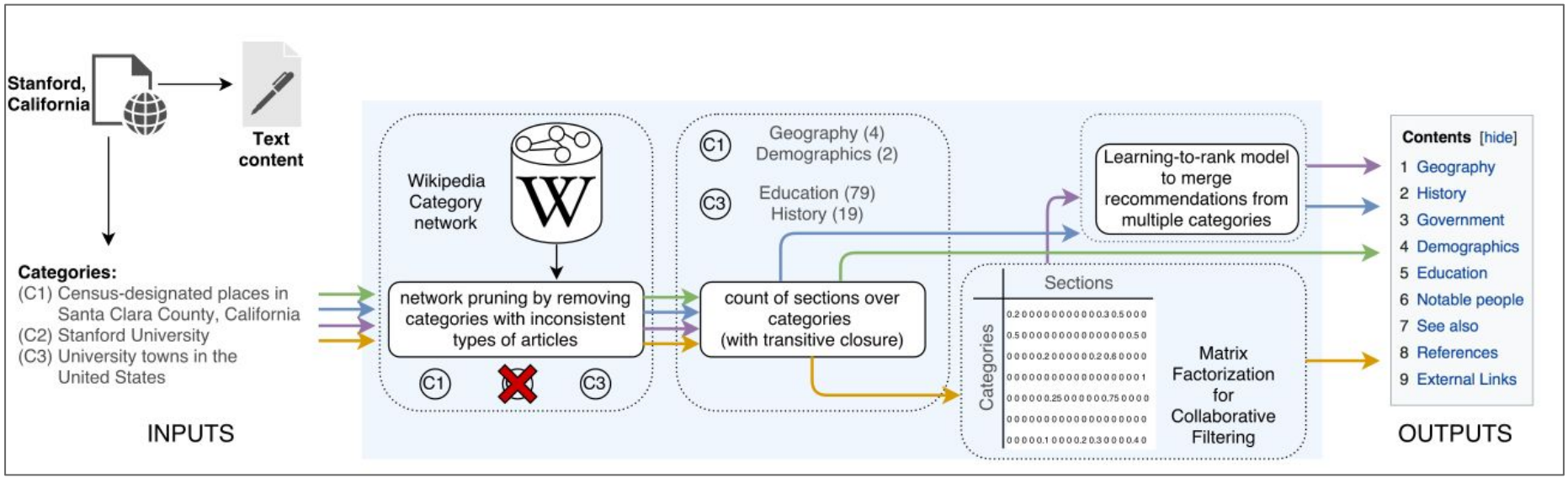
Experienced volunteer resources are scarce.

Editathon organizers and developers are manually creating article templates to help new contributors get started.

Why is this a hard problem?

- No existing template per article type defining which sections should exist
- No canonical way to define the topic of an article
- The problem needs to be solved for each of the 160+ Wikipedia language editions that are actively edited

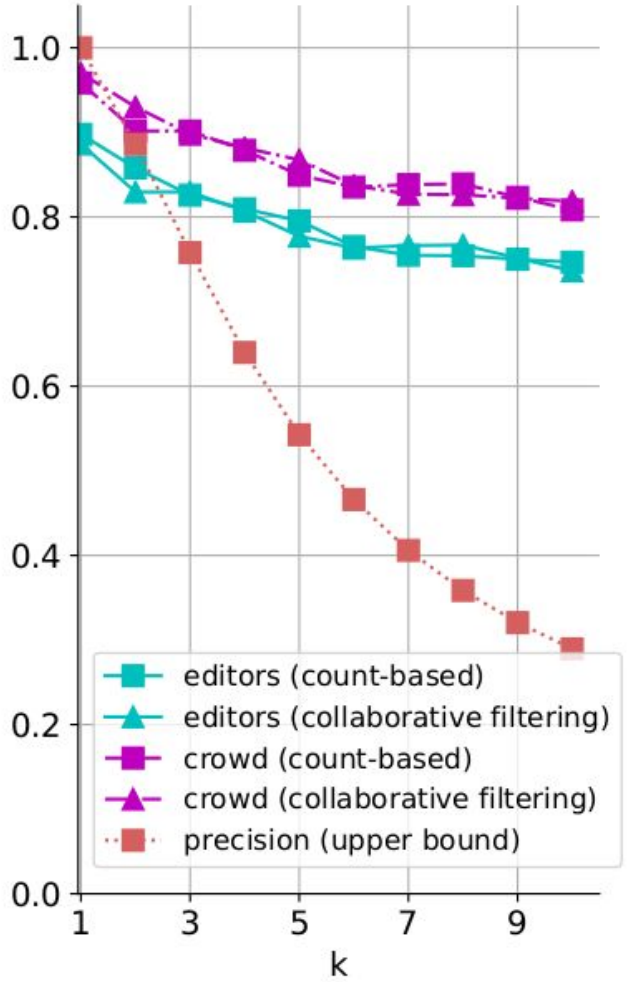
Our approach



meta:Research:Expanding_Wikipedia_articles_across_languages

Results

The section recommendation system we built has a very high performance when tested both with crowdworkers and experienced editors!



Technology Program 10

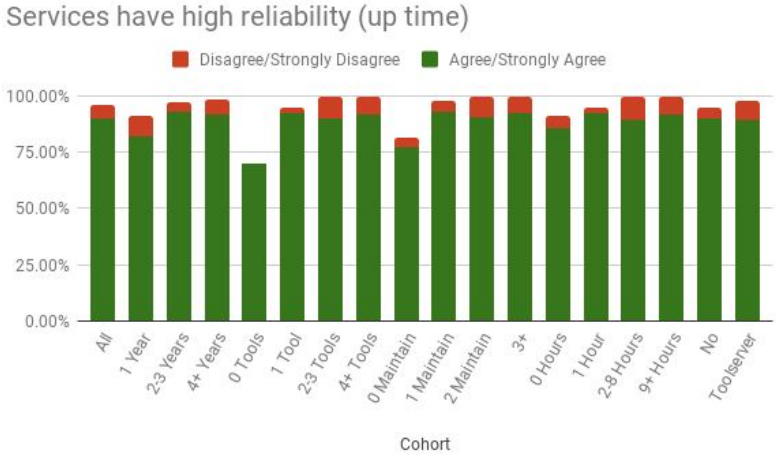
Public cloud services & support

Program Structure

Tech Community Support	TP10 Public Cloud Services & Support

Outcome 1 / Objective 2

Run annual Toolforge developer survey



WIKIMEDIA FOUNDATION

Outcome 2

The 'Labs, labs, labs' branding confusion is eliminated. Branding is separated, so that all of these are no longer referred to as just ‘Labs”: infrastructure as a service product, the platform as a service product, the team that manages those products, and the community that uses them to produce technical solutions.

Outcome 3

Promote Toolforge Tools and their maintainers within Wikimedia communities

WIKIMEDIA FOUNDATION

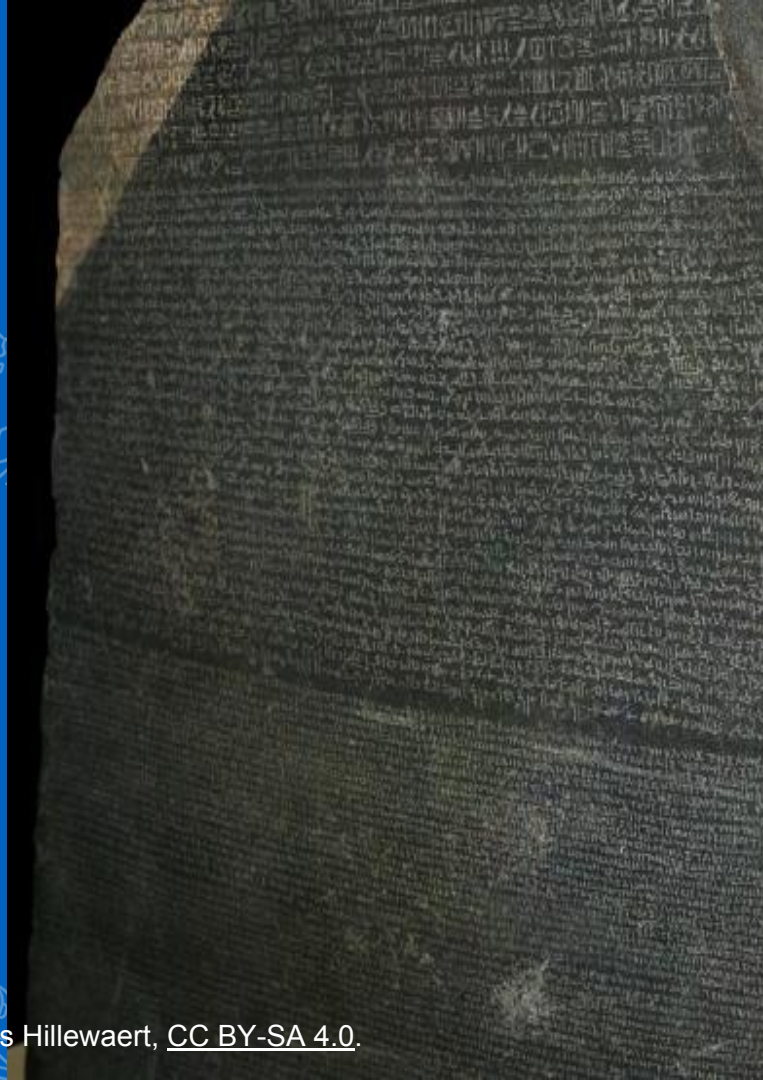


Photo by Hans Hillewaert, CC BY-SA 4.0.

Technology Program 11

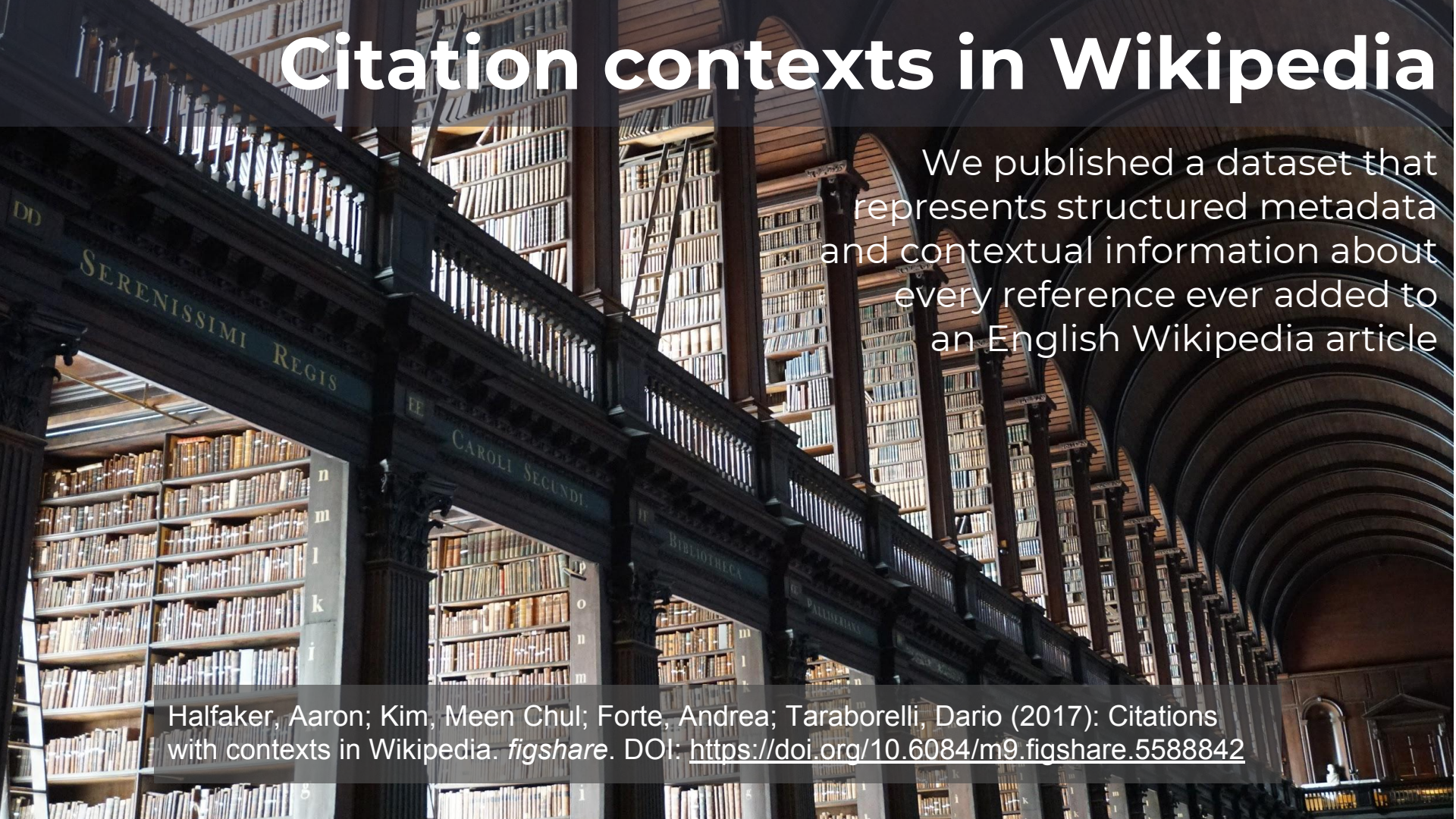
Improving citations across Wikimedia projects

Program Structure

Tech Community Support	TP11 Citations/Verifiability TP12 Growing Contributor Diversity X-CH Community Health/Anti-harassment

We will increase the verifiability of Wikimedia contents by conducting research aiming to improve how citations and sources are stored, accessed and vetted.

Citation contexts in Wikipedia

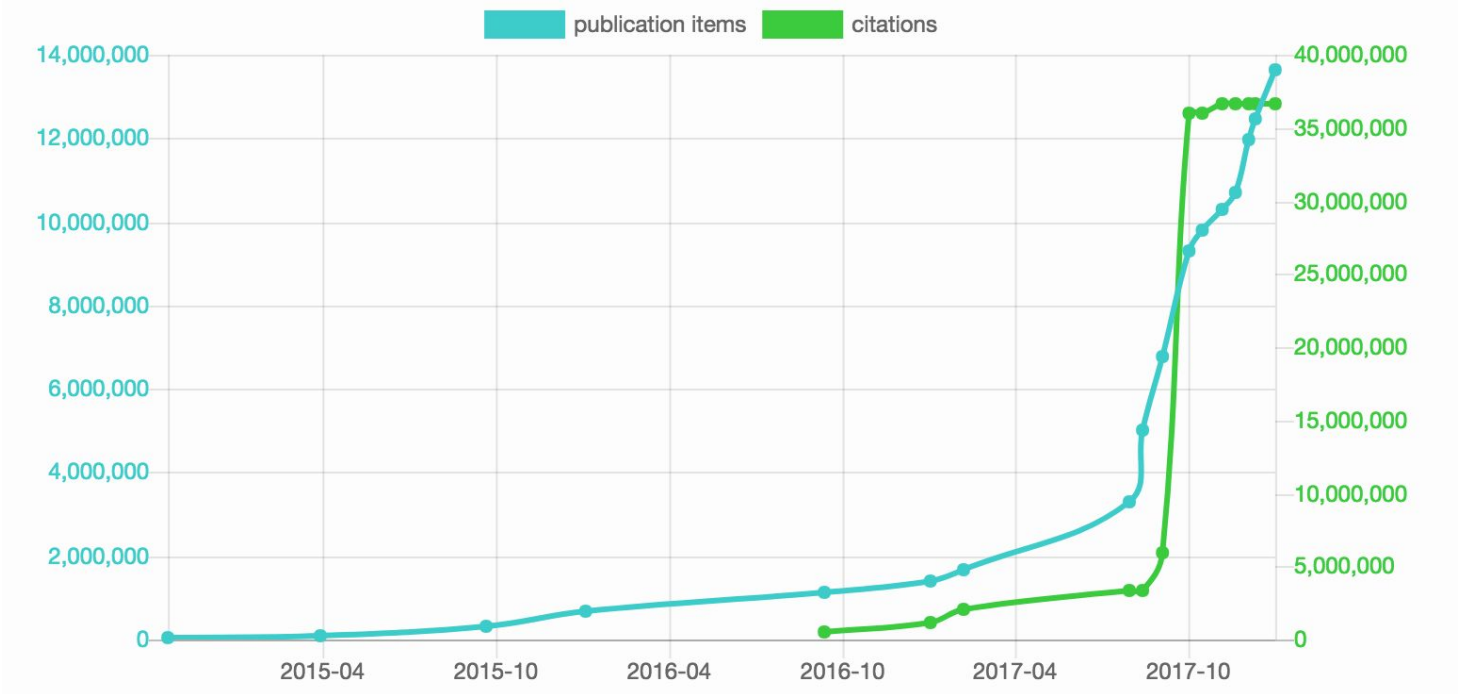


We published a dataset that represents structured metadata and contextual information about every reference ever added to an English Wikipedia article

Halfaker, Aaron; Kim, Meen Chul; Forte, Andrea; Taraborelli, Dario (2017): Citations with contexts in Wikipedia. *figshare*. DOI: https://doi.org/10.6084/m9.figshare.5588842

A growing knowledge base of sources in Wikidata

14 million items, 32% of all Wikidata



wikicite.org/statistics.html

WikiCite annual report

WikiCite 2017 group photo • Stephen Laporte, CC BY SA 2.0

About

WikiCite is an initiative aiming to build **a comprehensive knowledge base of sources**, to serve the sum of all human knowledge. In 2017, we convened nearly 100 attendees from 22 countries in Vienna for our annual event, to discuss progress, community needs and technical challenges towards this vision. This report examines the impact, key milestones, and reach the WikiCite community has achieved over the course of the past year.

WikiCite 2017 is generously supported by:

Alfred P. Sloan Foundation · Gordon and Betty Moore Foundation · Science Simons Foundation SANDBOX · Wikimedia Foundation · Wikimedia Österreich

Background

Wikipedia is the world's largest, most widely used online encyclopedia. It is free and open, a vast body of knowledge anyone can contribute to. Wikipedia's success lies in the neutrality and verifiability of the information it holds, its rigorous and transparent commitment to citation, fact-checking, and accuracy.

How does the Wikimedia movement empower individuals to assess reliable sources and arm them with quality information so they can make decisions based in facts? This question is relevant not only to Wikipedia users but to consumers of media around the globe.

Over the past decade, the Wikimedia movement has come together to answer that question. Efforts to design better ways to support sourcing have begun to coalesce around Wikidata – the free knowledge base that anyone can edit. With the creation of a rich, human-curated, and machine-readable knowledge base of sources, the WikiCite initiative is crowdsourcing the process of vetting information and its provenance.

Citation needed • Dario Taraborelli, CC0

commons:File:WikiCite_2017_report.pdf