TECHNOLOGY DEPT

Community Support Programs

May 2018 quarterly check-in

for work done in Q3 FY2017/18

WIKIMEDIA
FOUNDATION

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

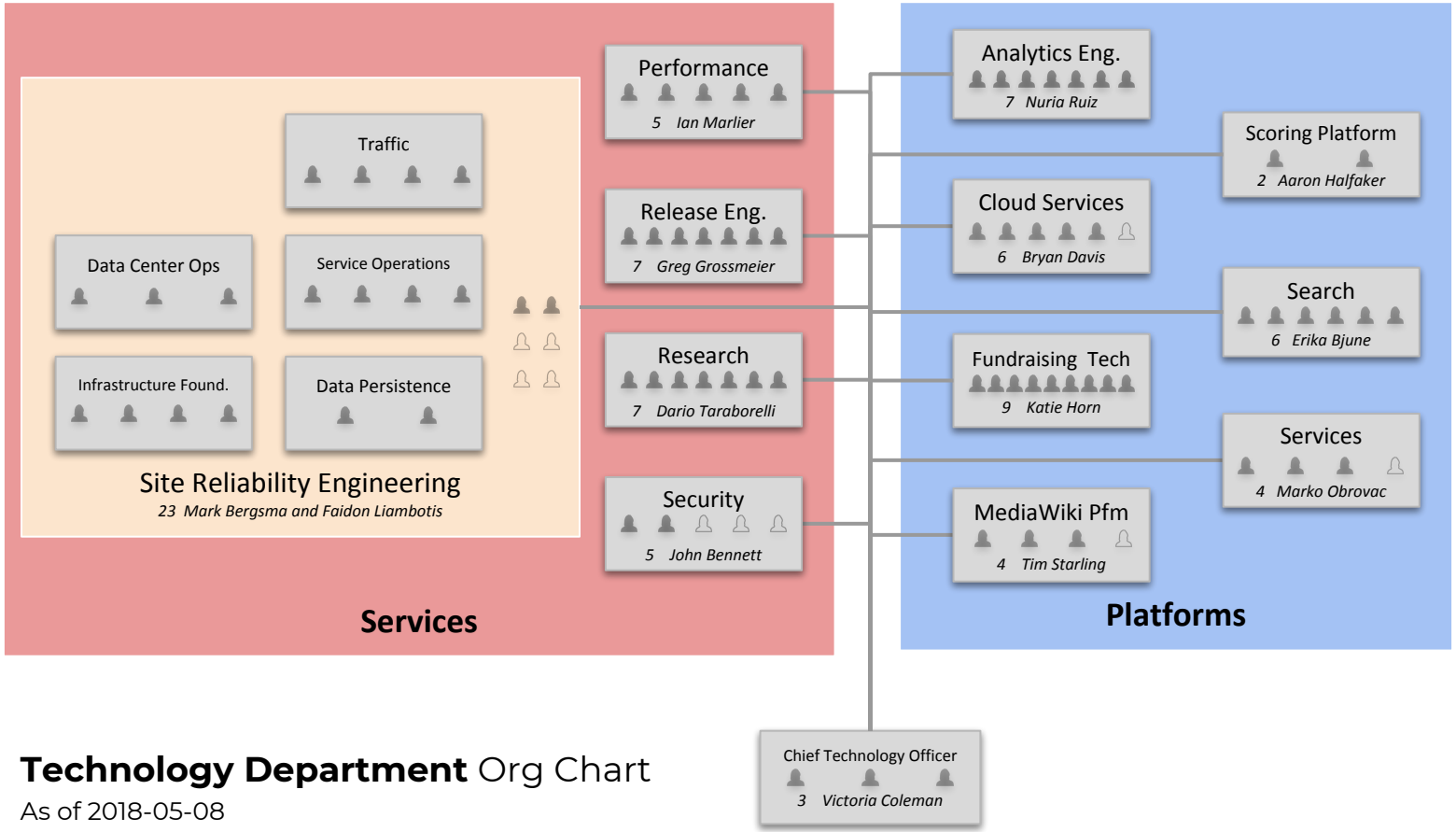
What are other people asking from us?

Improvement

What could we do to improve our offering?

Maintenance

What will sustain and improve our delivery?



Technology Department Org Chart

As of 2018-05-08

Includes filled and vacant Reqs only

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 11: Improving citations across Wikimedia projects

Programs covered in other presentations:

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

No goals this quarter — Program 10: Public cloud services and support

Technology Program 4

Technical community building

Program Structure

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 1:

Form a documentation Special Interest Group

Technical Document Re-working Group

WIKIMEDIA
FOUNDATION

Outcome 1 / Objective 3:

Increase community awareness of volunteer developed tools

Add information

Adds an {{Information}} template to a file on Commons.

By Magnus Manske (source available)

commons information template

GLAMorous 2

A tool to keep track usage and views of Commons images on other projects.

By Magnus Manske (source available)

glam files images views commons

wdtaxonomy

command line tool to extract taxonomies from Wikidata

By Jakob Voß (source available)

wikidata classes

Examples of map layers

A collection of various map layers and tile services as used within Wikimedia, Tools etc. This is a fork of https://github.com/leaflet-extras /leaflet-providers

By Derk-Jan Hartman (source available)

demo leaflet maps

Section Links

This tool shows wikilinks to inexistent section titles from or to a page.

By Pietrodn (source available)

tools sections articles wikilinks links

Wiki Loves Monuments UK 2014

2014 U.K Wiki Loves Monuments interface

By Magnus Manske

wlm wiki loves monuments monuments wikidata commons

Wiki ViewStats

Pageview statistics for all Wikimedia wikis. Features: TOP-Lists per category, Wildcard search, disambiguation helper, cross-wiki pageviews per page and more.

By Hedonil

statistics pageviews hitcount

Wikipedia Cite-o-Meter

Find citations by publisher in the top 100 Wikipedias

By Dario Taraborelli (source available)

wikipedia cite citation

Contributors

Creates a list of contributors to a given article on a given project in wikitext.

contributor page history

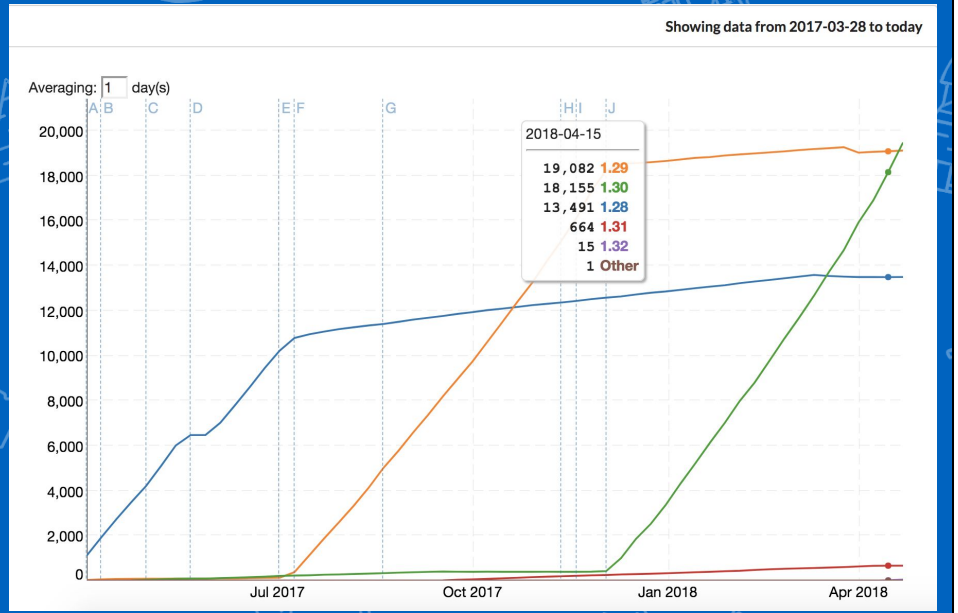
Outcome 3/ Objective 1:

Establish ongoing channels of communication with third-party developers

Participated in:

- Enterprise MediaWiki Conference (EMWCon)
  - Helped organize
  - 6 staff attended
  - 4 staff presented
  - 5 participated in Create Camp
  - Trip report on office wiki
- Enterprise MediaWiki Slack team (NASA) transitioned to FOSS Riot.im (T184606)
  - 9 rooms, 31 members, and climbing
- US Federal Government MediaWiki group
  - Voluntary Product Accessibility Template
- MediaWiki Stakeholders Group

How is MediaWiki being used?



Outcome 3/Objective 2:

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

Pingback:

- MediaWiki version, database type, PHP version, operating system, machine architecture, processor family, web server, memory limit
- Gathering data since March 2017
- Q3: Made aggregate data available at https://pingback.wmflabs.org/
- Q3: Added monthly heartbeat ping

Outcome 4 / Objective 1:
Organize Wiki Workshop 2018

In collaboration with our co-organizers at EPFL and Stanford:

- 6 keynote speakers
- 14 PC members
- 21 accepted papers
- More than 100 attendees

The workshop was held at The Web Conference in Lyon, France in Q4

http://wikiworkshop.org/2018/


Image by Nathan Dumlao

Developer Summit: *movement strategy*

8 topics covered in 2 days:

Knowledge as a Service
Supporting Third-Party Use of MediaWiki
Evolving the MediaWiki Architecture
Next Steps for Languages and Cross Project Collaboration
Advancing the Contributor Experience
Growing the MediaWiki Technical Community
Embracing Open Source Software
Research, Analytics, and Machine Learning

Open, thoughtful, and impactful conversations were had in support of the vision and strategy of the Wikimedia movement.

Image by Karly Jones

WM Technical Conference: *platform evolution*

Taking the time to make informed decisions in the evolution of our platform and influence the 3 - 5 year strategic goals and roadmap.

*"Empower the Wikimedia Foundation to accomplish its goals of **Knowledge Equity** and **Knowledge as a Service** by evolving and investing in our technology stack to improve its flexibility, maintainability, and sustainability"*

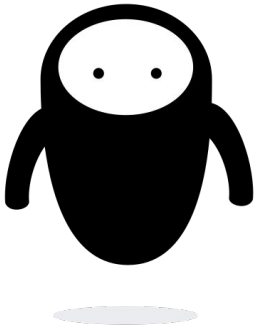
Technology Program 5

Scoring Platform (ORES)

Program Structure

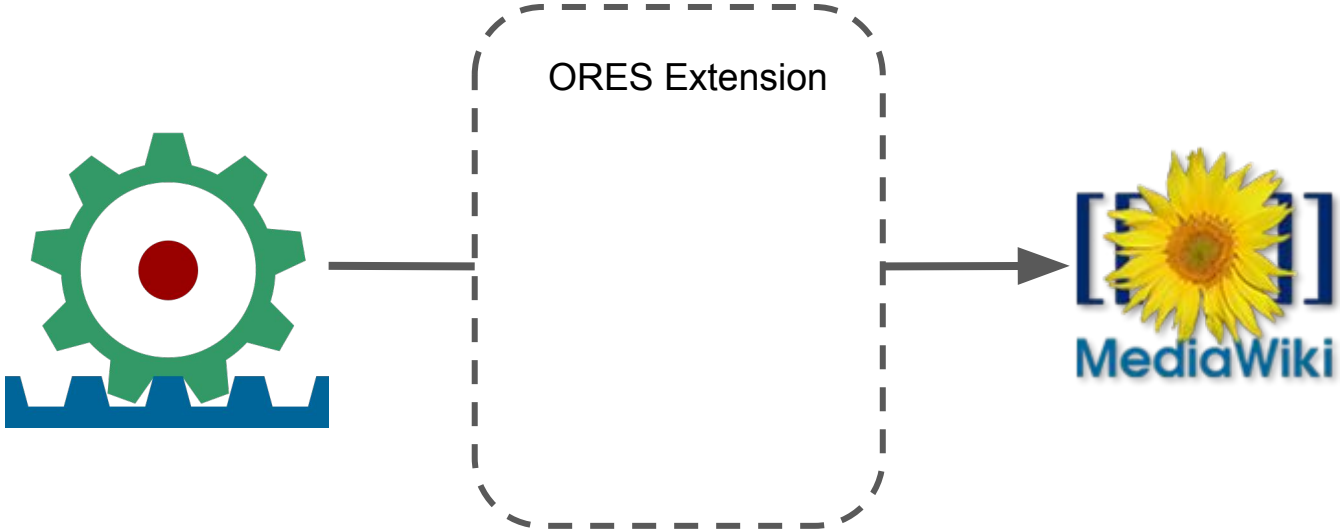
Community Support	TP5 Scoring Platform (ORES) TP9 Growing Wikipedia Across Languages TP7 Smart Tools for Better Data

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

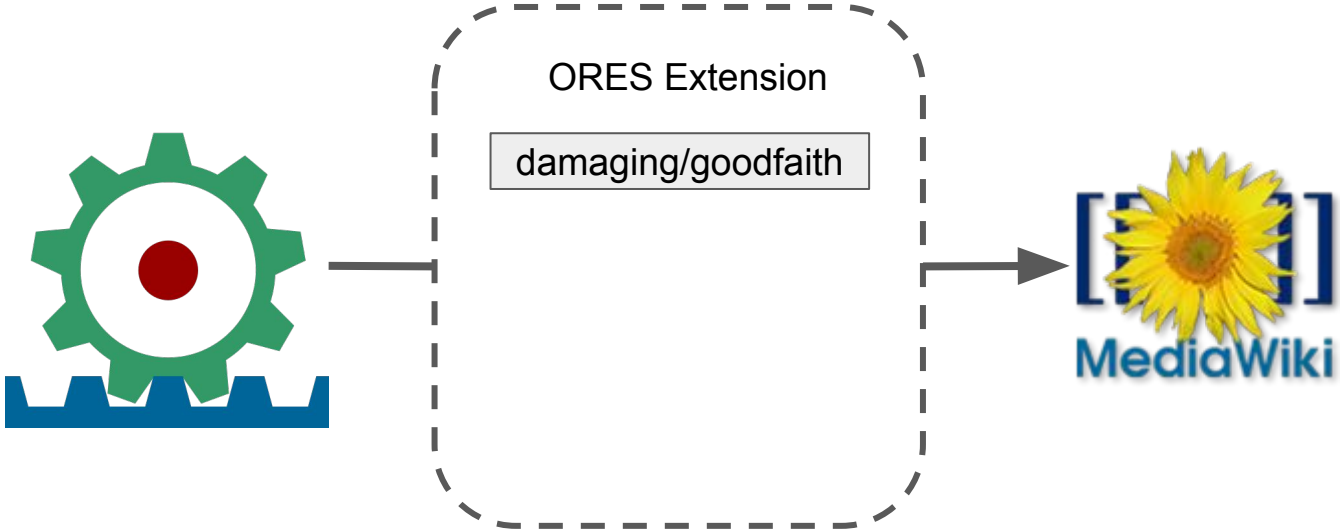


SCORING PLATFORM TEAM

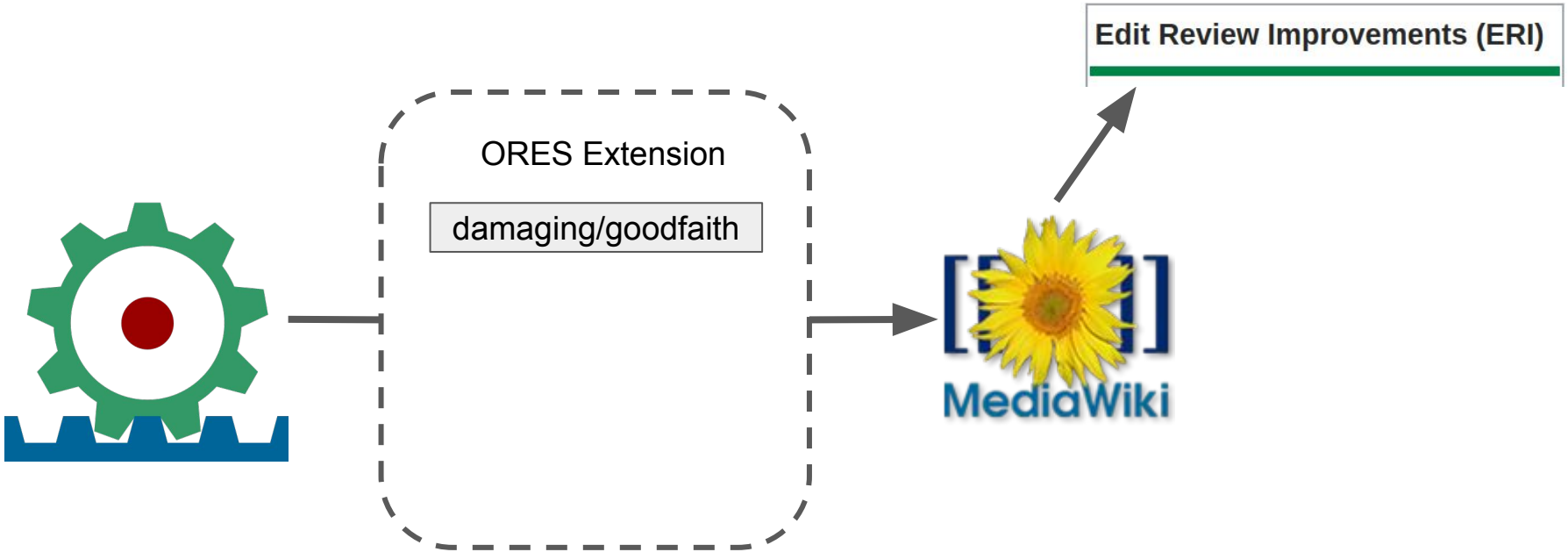
Expanding ORES integration in MediaWiki



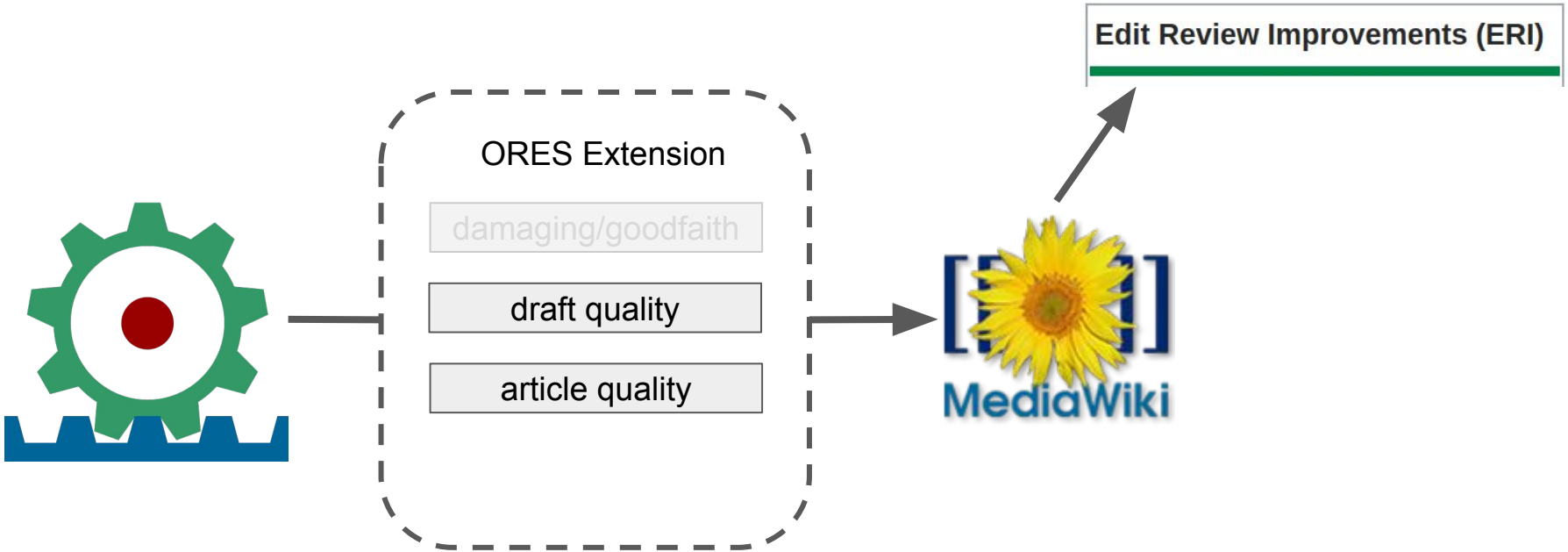
Expanding ORES integration in MediaWiki



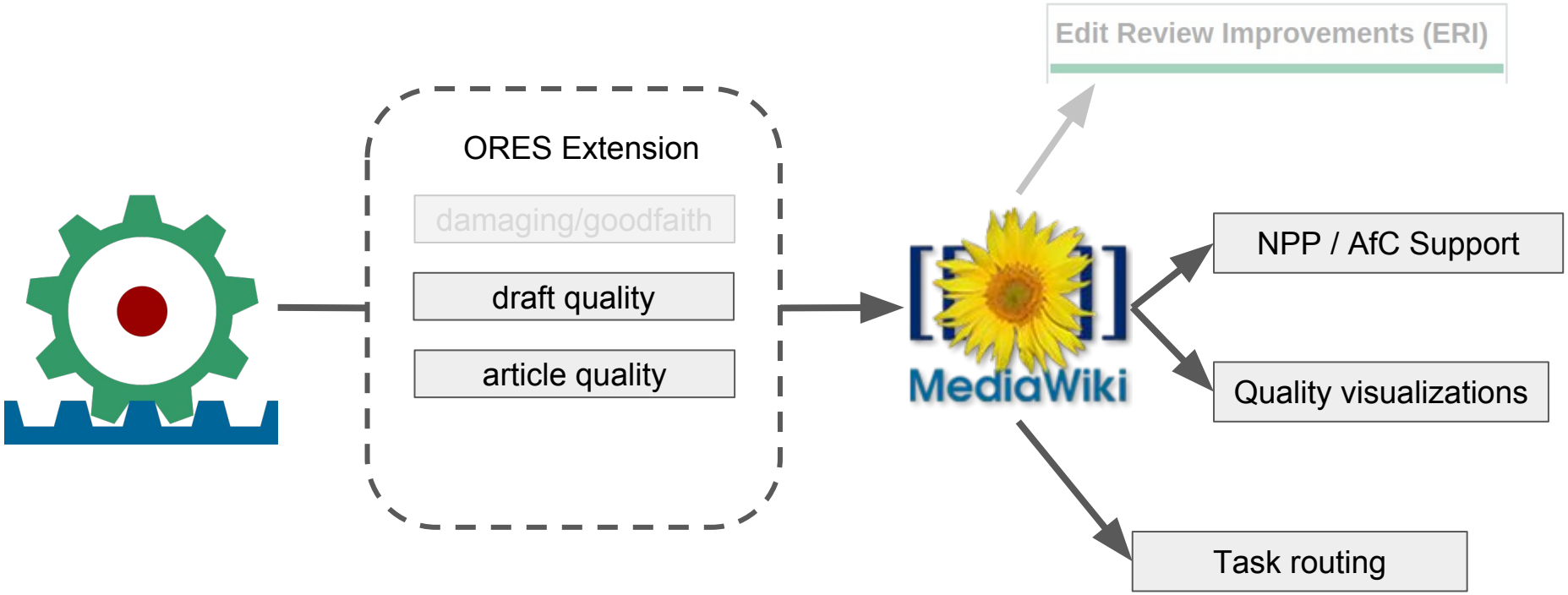
Expanding ORES integration in MediaWiki



Expanding ORES integration in MediaWiki

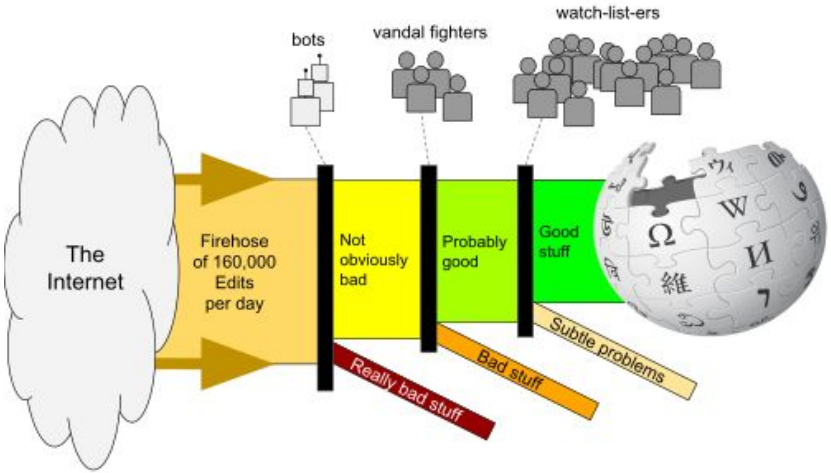


Expanding ORES integration in MediaWiki

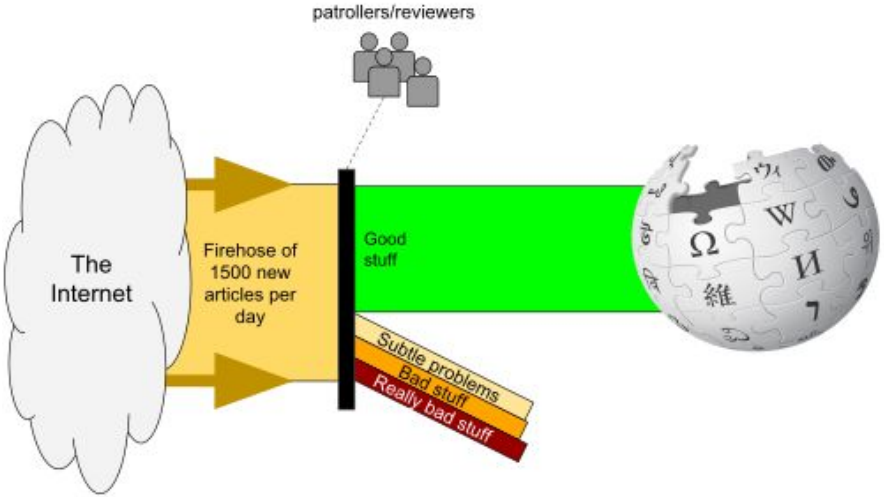


Draft topic -- **expanding new article review**

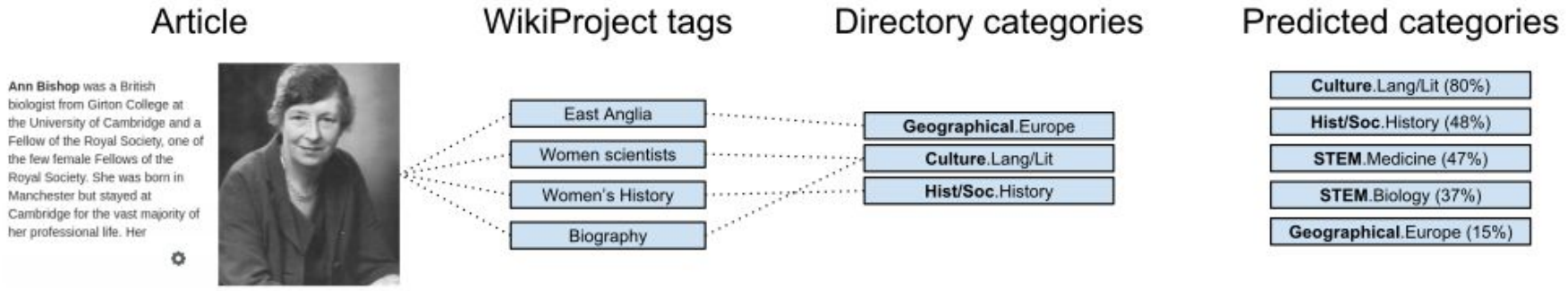
New edit review



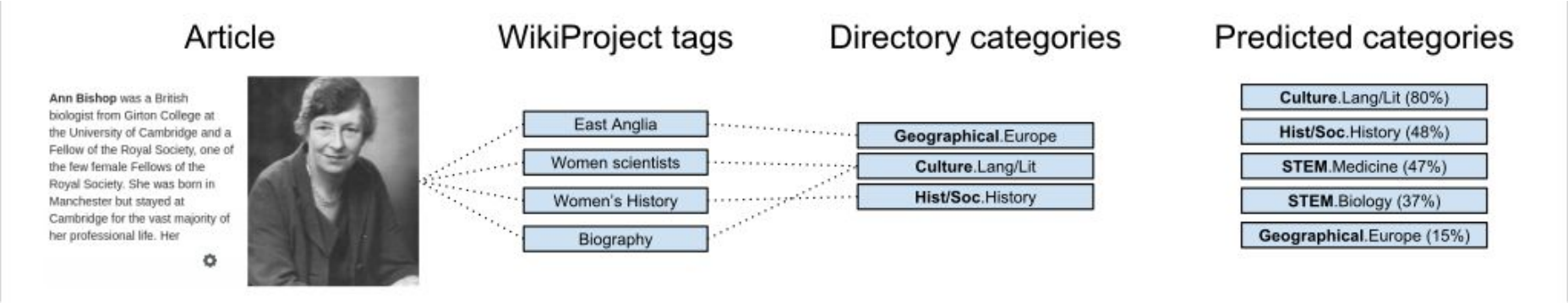
New article review



Draft topic -- **expanding new article review**

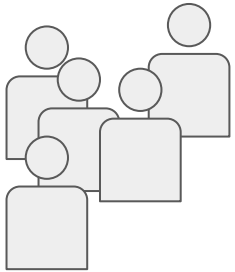


Draft topic -- **expanding new article review**



User:Rosiestep

+



New draft creators (creating drafts about *women writers*)

=

- Less work for patrollers
- More good content sticks
- Better socialization for newcomers
- More contributors joining WikiProjects

Status

1. Develop draft topic model
2. Increase fitness
3. Implement feature extraction strategy
4. Develop deployment strategy
5. Deploy large assets with ORES
6. Deploy the draft topic model

Status

1. Develop draft topic model ✓
2. Increase fitness ✓
3. Implement feature extraction strategy ✓
4. Develop deployment strategy ✓
5. Deploy large assets with ORES ✘
6. Deploy the draft topic model **??? → Q4**

Status

1. Develop draft topic model ✓
2. Increase fitness ✓
3. Implement feature extraction strategy ✓
4. Develop deployment strategy ✓
5. Deploy large assets with ORES ✘
6. Deploy the draft topic model **??? → Q4**

[[:mw:Wikimedia Research/Showcase]]

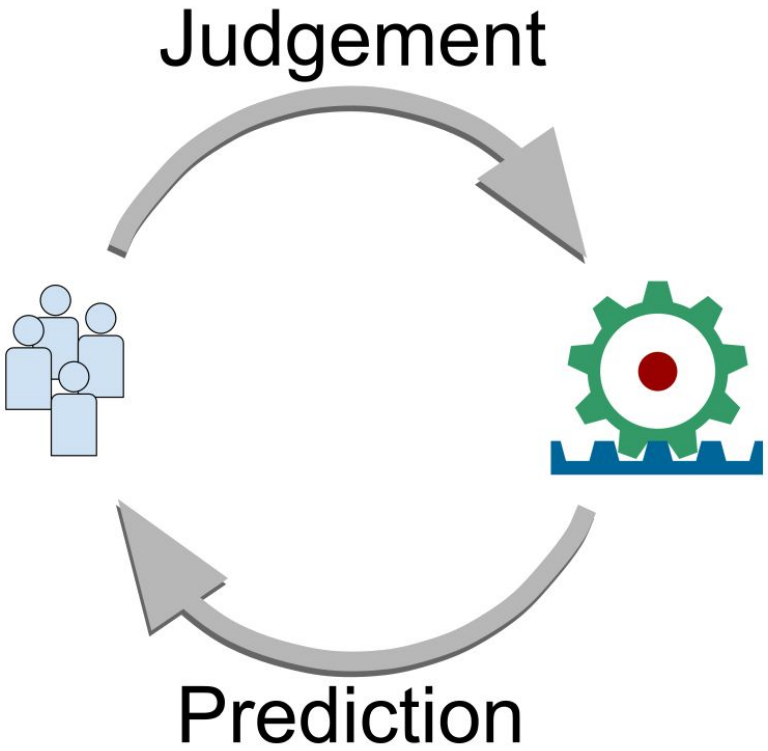
Wikimania Talk Proposal (Accepted)

JADE (MVP in Beta cluster)

JADE:Diff/376901

From Wikipedia


```
{
  "entity": {
    "type": "diff",
    "rev_id": 376901
  },
  "scores": {
    "schema": {
      "spec": "https://phabricator.wikimedia.org/diffusion/
               view=raw",
      "name": "damaging"
    },
    "data": {
      "damaging": true
    }
    "schema": {
      "spec": "https://phabricator.wikimedia.org/diffusion/
               view=raw",
      "name": "goodfaith"
    },
    "data": {
      "goodfaith": false
    }
```

Other notable milestones

Wikimania proposals submitted

- Using Artificial Intelligence to keep Wikipedia open
- JADE -- Support for Auditing AIs
- Draft review routing via topic modeling

New advanced support for

- Latvian Wikipedia
- Hungarian Wikipedia
- Arabic Wikipedia
- Catalan Wikipedia

Supporting anti-harassment

- Elizabeth Whittaker (intern) started working with us to model Civility

ORES Support Checklist
(now with automation!)

Last updated on 03 May 2018 12:01:39 UTC

Wiki	edit quality			article quality			
	Basic support	Advanced support		wp10		draftquality	
		Labeling campaign	model	Labeling campaign	model	Labeling campaign	model
arwiki	n/a	99%	✓				
azwiki		0%					
bawiki		0%					
bnwiki	✓	3%					
bnwikisource		0%					
cawiki	n/a	100%	✓				
cswiki	n/a		✓				
dewiki	✓	11%					
elwiki	✓						
enwiki	n/a	50%	✓		✓		✓
enwiktionary	✓	0%					
eswiki	n/a		✓				
eswikibooks	n/a		✓				
eswikiquote	✓	5%					
etwiki	n/a		✓				
fawiki	n/a		✓	10%			
fiwiki	n/a	51%	✓				
frwiki	n/a		✓		✓		
frwikisource							
hewiki	n/a		✓				
hrwiki	✓	17%					
huwiki	n/a	100%	✓				
idwiki	✓	4%					
iswiki	✓						
itwiki	✓	20%					
jawiki		3%					
kowiki	✓	54%					
lvwiki	n/a	99%	✓				
nlwiki	n/a		✓				
nowiki	✓	41%					
plwiki	n/a		✓				
ptwiki	n/a		✓				
rowiki	n/a		✓				
ruwiki	n/a		✓		✓		
simplewiki	n/a	54%	✓		✓		✓
sqwiki	n/a		✓				
srwiki		68%					
svwiki	n/a		✓				
tawiki	✓						
testwiki	✓		✓				
trwiki	n/a	8%	✓		✓		
ukwiki	✓	18%					
urwiki		3%					
viwiki	✓	3%					
wikidatawiki	n/a		✓				
zhwiki		100%					

https://phabricator.wikimedia.org/phame/post/view/104/status_update_may_2_2018/

Status Update (May 2, 2018)

Public

Actions

Highlights

- We've started work on JADE in earnest, and the prototype is deployed to the beta cluster where it's available for testing and tool development.
- Draft topic prerequisites are mostly falling into place, so we should be able to get the initial model deployed this month.
- New, dynamic ORES support table shows up-to-date information about our progress for each wiki: https://tools.wmflabs.org/ores-support-checklist/
- ORES is served from its own cluster, which gave us a tremendous benefit in both performance and stability.
- More ORES support for Arabic, Bengali, Catalan, Hungarian, Latvian, Swedish Wikipedia

Outreach

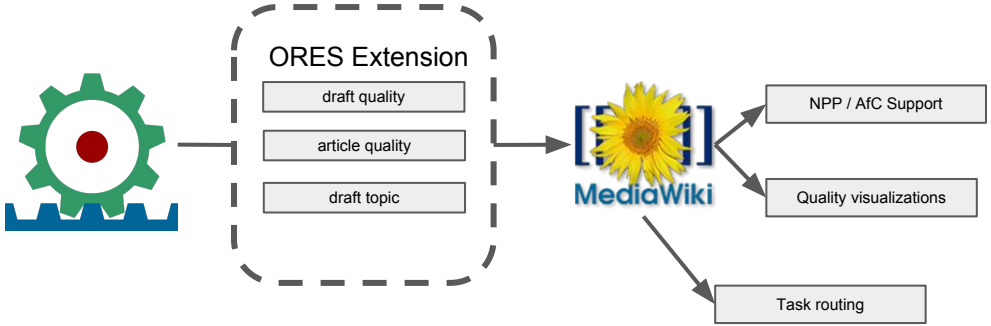
~~T121719: [Epic] Write paper about ORES as a socio-technical probe~~
~~T188123: Present about draft topic model at Wikimedia Research Showcase.~~
~~T188124: Build slide deck about AI at Wikimedia for Policy People~~
~~T190464: Discuss surfacing ORES for AFC/NPP~~

Draft topic

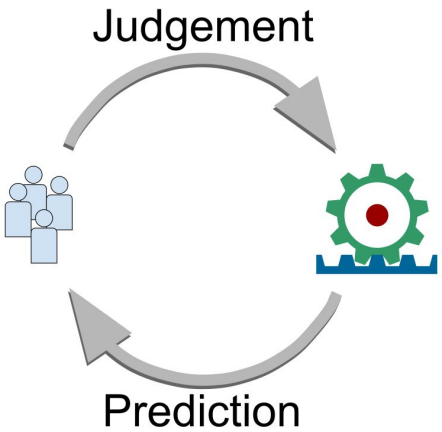
~~T123327: Train/test draft topic model (new article routing AI)~~
~~T185147: Host Google-News-word2vec.bin publicly~~
~~T185896: OneVsRest Classification for revscoring~~
~~T188445: Implement word2vec featurevector in revscoring~~
~~T189364: Investigate word2vec memory issues with multiprocessing~~

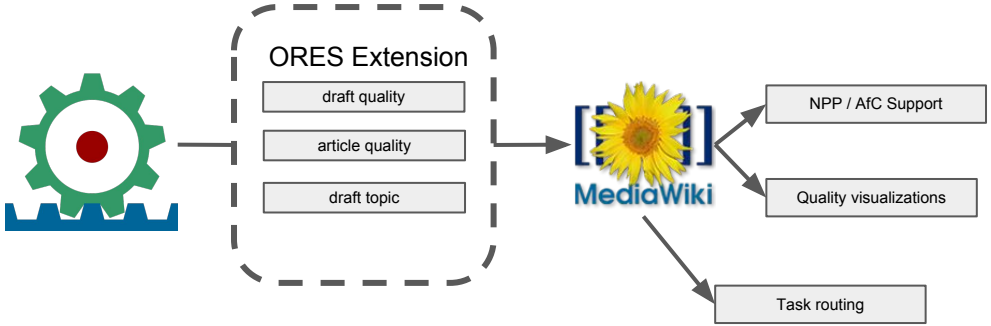
Q4

Q4


ORES Extension
draft quality
article quality
draft topic
MediaWiki
NPP / AfC Support
Quality visualizations
Task routing

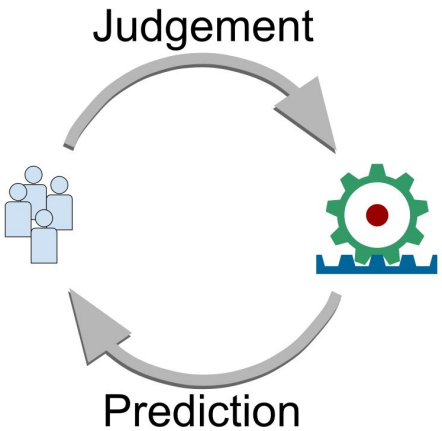
Q4



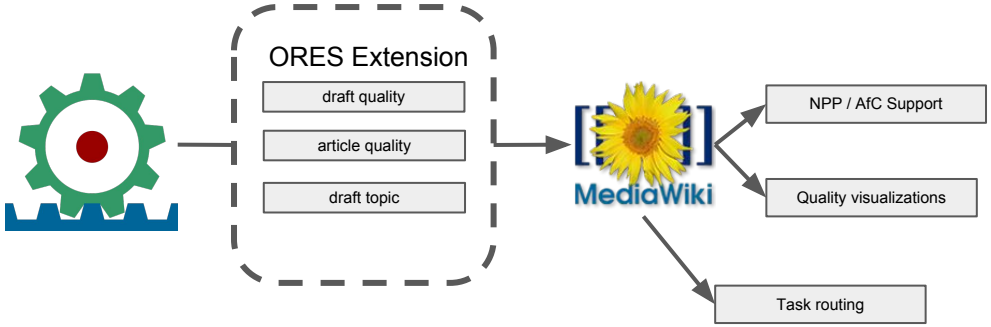


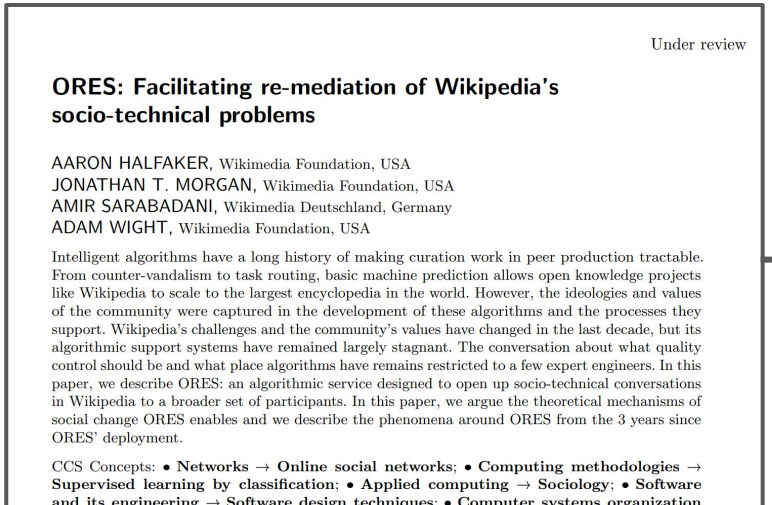
ORES/JADE Integration

Q4



ORES/JADE Integration





Under review

ORES: Facilitating re-mediation of Wikipedia's socio-technical problems

AARON HALFAKER, Wikimedia Foundation, USA
JONATHAN T. MORGAN, Wikimedia Foundation, USA
AMIR SARABADANI, Wikimedia Deutschland, Germany
ADAM WIGHT, Wikimedia Foundation, USA

Intelligent algorithms have a long history of making curation work in peer production tractable. From counter-vandalism to task routing, basic machine prediction allows open knowledge projects like Wikipedia to scale to the largest encyclopedia in the world. However, the ideologies and values of the community were captured in the development of these algorithms and the processes they support. Wikipedia's challenges and the community's values have changed in the last decade, but its algorithmic support systems have remained largely stagnant. The conversation about what quality control should be and what place algorithms have remains restricted to a few expert engineers. In this paper, we describe ORES: an algorithmic service designed to open up socio-technical conversations in Wikipedia to a broader set of participants. In this paper, we argue the theoretical mechanisms of social change ORES enables and we describe the phenomena around ORES from the 3 years since ORES' deployment.

CCS Concepts: • **Networks** → **Online social networks**; • **Computing methodologies** → **Supervised learning by classification**; • **Applied computing** → **Sociology**; • **Software and its engineering** → **Software design techniques**; • **Computer systems organization**

AI @ WMF
(Strategy)
3-5 year

Wikimedia Scoring Platform team

Welcome to the home of the **Wikimedia Scoring Platform team**. For our past work as an ad-hoc, volunteer project, see m:Research:Revision scoring as a service. As of July, 2017, we're an officially funded team operating within the Technology Department at the Wikimedia Foundation.

Contents [show]

Team [edit]

Aaron Halfaker
Principal Research Scientist
Team Lead

Amir Sarabadani
Software Engineer (WMDE)

Adam Wight
Software Engineer
(WMF)

Ewhit
Research Intern (WMF)

Former

- とある白い猫 (IEG)
- Arthur Tilley (IEG)
- He7d3r (IEG/Volunteer)
- Yuvipanda (Volunteer)
- Sumit (Volunteer)

Wikimedia engineering activity

Scoring Platform

We build scoring services that enable advanced wiki tools

Group: Technology

Team: **Staff**

- Aaron Halfaker
- Adam Wight
- Amir Sarabadani

Volunteers

- Sumit

Management: Aaron Halfaker

SCORING PLATFORM TEAM

Wikimedia Scoring Platform team logo

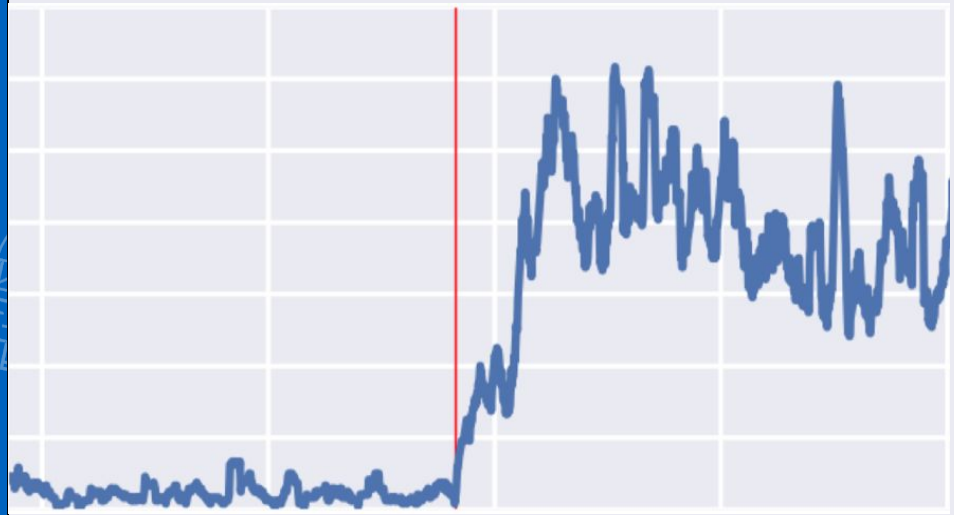
Technology Program 7

Smart tools for better data

Program Structure

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.



New Private Dataset: GeoEditors

Editor data per **activity level,** per **project,** per **country** is now available internally at http://superset.wikimedia.org

Geoeditors Monthly - Editor Data Per Country

Geoeditors Filter

Since 2018-02-01

Until 2018-03-01

wiki_db

× eswiki

activity_level

Select [activity_level]

users_are_anonymous

Select [users_are_anonymous]

Geoeditors - Map

Geoeditors - Table

country_code	SUM(distinct_editors)
ES	22.8k
MX	14.9k
CO	9.41k
AR	8.43k
CL	4.73k
PE	3.42k
VE	3.20k
US	1.97k
EC	1.72k
DO	1.40k
UY	1.29k
GT	936
CR	707
BO	685
SV	569
HN	567
PY	549
BR	488
DE	476
FR	471
IT	420
GB	362
PA	317
NI	306
PR	252
CA	180

Geoeditors Monthly - Editor Data Per Country

Geoeditors Filter

Since 2018-02-01

Until 2018-03-01

wiki_db

× eswiki

activity_level

× 100 or more

users_are_anonymous

Select [users_are_anonymous]

Geoeditors - Map

Geoeditors - Table

country_code	SUM(distinct_editors)
ES	243
AR	105
MX	63.0
CL	57.0
CO	38.0
PE	35.0
VE	23.0
UY	13.0
EC	10.0
BR	9.00
CR	8.00
PA	8.00
DE	6.00
US	6.00
BO	4.00
NL	4.00
DO	3.00
FR	3.00
GB	3.00
GT	3.00
PY	3.00
CH	2.00
HN	2.00
IE	2.00
NI	2.00
NO	2.00

Geoeditors - Editors Per Wiki (top 50)

New Private Dataset: GeoEditors

We want to make this data public,

will start working with legal in Q1.

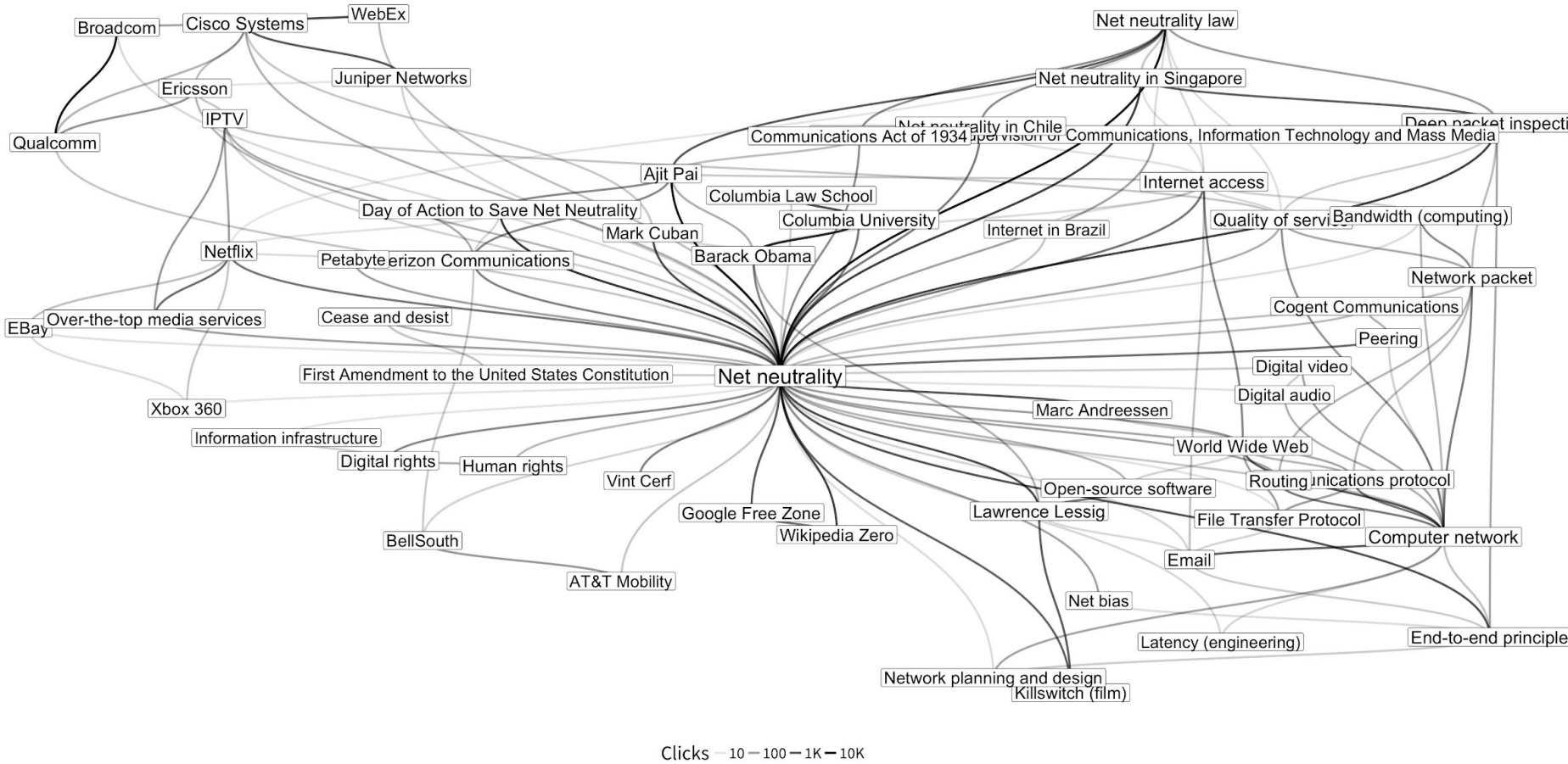
New Public Dataset: ClickStream

The dataset represents

-*in* ***aggregate***- how **readers** reach

a Wikipedia article and **navigate**

to the next.


Broadcom
Cisco Systems
WebEx
Juniper Networks
Ericsson
IPTV
Qualcomm
Net neutrality law
Net neutrality in Singapore
Net neutrality in Chile
Communications Act of 1934
Communications, Information Technology and Mass Media
Ajit Pai
Columbia Law School
Columbia University
Internet access
Internet in Brazil
Day of Action to Save Net Neutrality
Mark Cuban
Barack Obama
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
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

High Volume Eventlogging in Hadoop.

Measuring page previews and others.

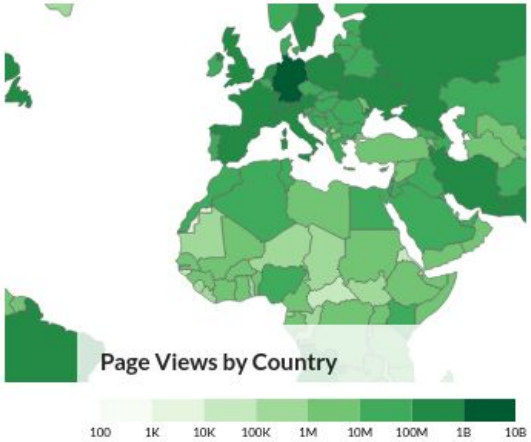
Wikistats 2.

Pageviews per country and mobile friendly UI.

Wikimedia Statistics

Page Views by Country *Monthly*

All wikis



Countries where this project is visited the most. Those countries with less than 100 views are not reported and are blank in the map.. More info about this metric.

Wikimedia Statistics

Dashboard

Contributing

Reading

Content

Total Page Views

681M

April ↓ *-6.85 % month over month*

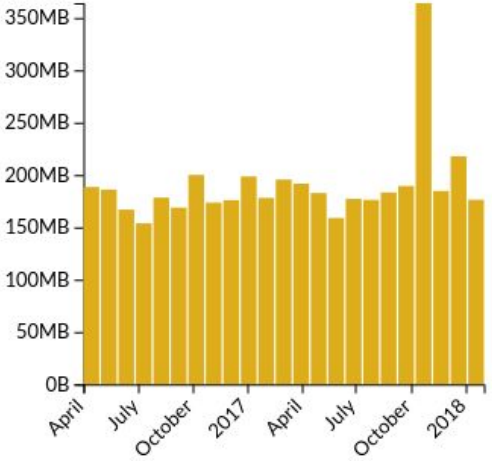
9B ↑ *0.40 % year over year*

Year total (2017)

Wikimedia Statistics

Net bytes difference *Monthly*

Wikipedia – French



Average: 190MB ↓ -6.42% over this time range.

The sum of the differences in bytes made by each edit (or revision), including edits on redirects. More info about this metric.

Wikistats 2

Still a lot of work to do in the backend infrastructure for wikistats data; our next round of changes will not be visible on UI.

Complete storage scaling migration

Services **fully migrated** the RESTBase production cluster **to Cassandra 3** and migrated all of the use cases to the new storage design.

Drill-down-type dashboards have also been created and they will be shared with other Cassandra clusters (AQS, Maps).

After switching we encountered performance problems, likely related to SSDs we have been using.

- Investigation is ongoing in Q4.

Outcome 3 /
Objective 1:

Provide reliable and available access to Wikimedia database dumps

SixHardDriveFormFactors.jpg, CC-BY-SA 3.0

Technology Program 9

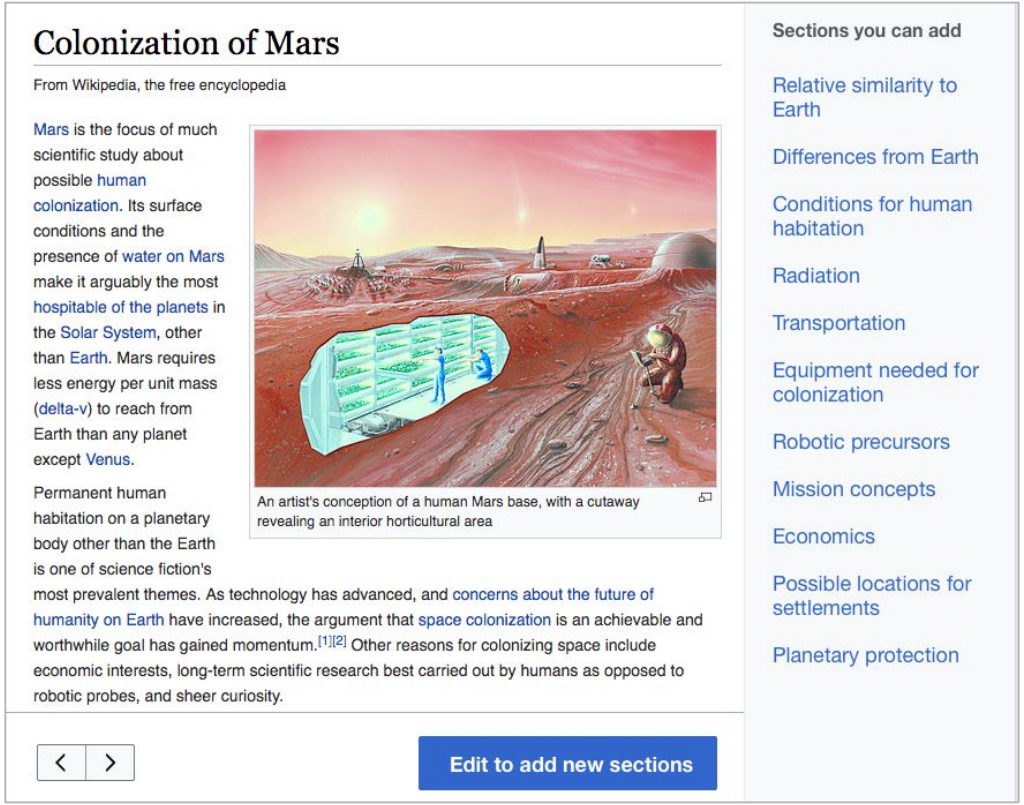
Growing Wikipedia across languages

Program Structure

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

The problem we're trying to solve

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



Colonization of Mars

From Wikipedia, the free encyclopedia

Mars is the focus of much scientific study about possible human colonization. Its surface conditions and the presence of water on Mars make it arguably the most hospitable of the planets in the Solar System, other than Earth. Mars requires less energy per unit mass (delta-v) to reach from Earth than any planet except Venus.

An artist's conception of a human Mars base, with a cutaway revealing an interior horticultural area

Permanent human habitation on a planetary body other than the Earth is one of science fiction's most prevalent themes. As technology has advanced, and concerns about the future of humanity on Earth have increased, the argument that space colonization is an achievable and worthwhile goal has gained momentum.[1][2] Other reasons for colonizing space include economic interests, long-term scientific research best carried out by humans as opposed to robotic probes, and sheer curiosity.

Sections you can add

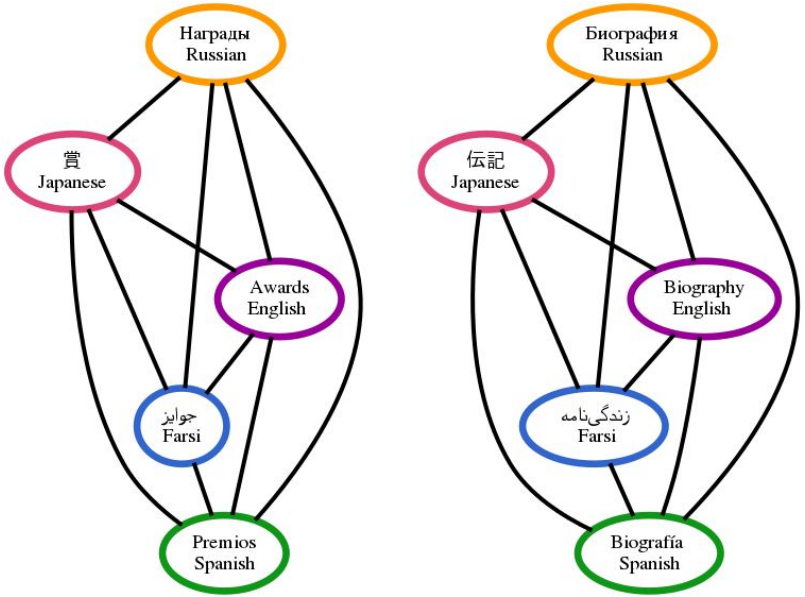
- Relative similarity to Earth
- Differences from Earth
- Conditions for human habitation
- Radiation
- Transportation
- Equipment needed for colonization
- Robotic precursors
- Mission concepts
- Economics
- Possible locations for settlements
- Planetary protection

Edit to add new sections

Section translation / synonym classifiers

1) Building a ground truth test set

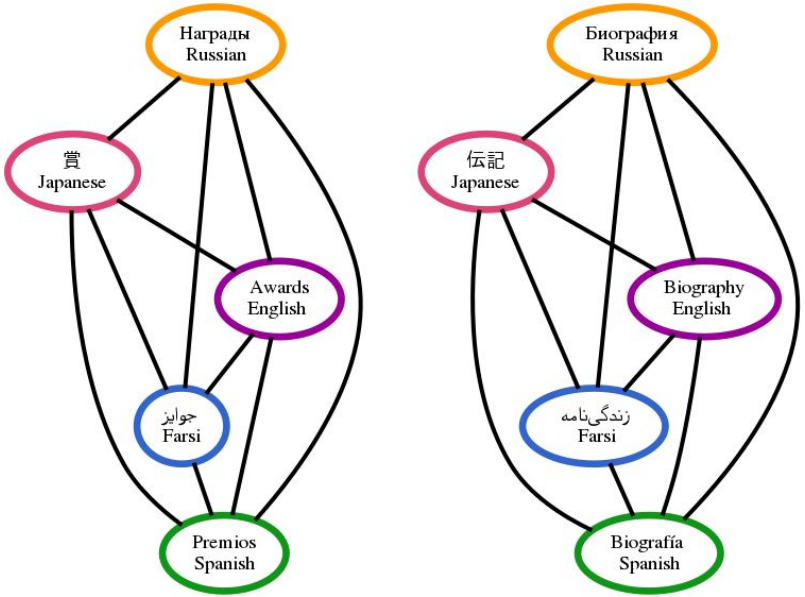
- Defined 6 languages to work with from different families and scripts: **ar, es, en, fr, ja, ru**.
- Generated a labeling set for these languages
- Developed a method to find bilingual editors.
- We have already obtained synonym labels in English, Russian and French, and the labeling task is still in process in Q4.



Section translation / synonym classifiers

2) Designing and testing the models

- Developed a first **section translation classifier** based on Wikidata interlingual links and cross-lingual word embeddings.
- Tested a **section synonym classifier** in English and Russian, with promising results (over 94% accuracy) in cross-lingual experiments (training in one language and testing in another).

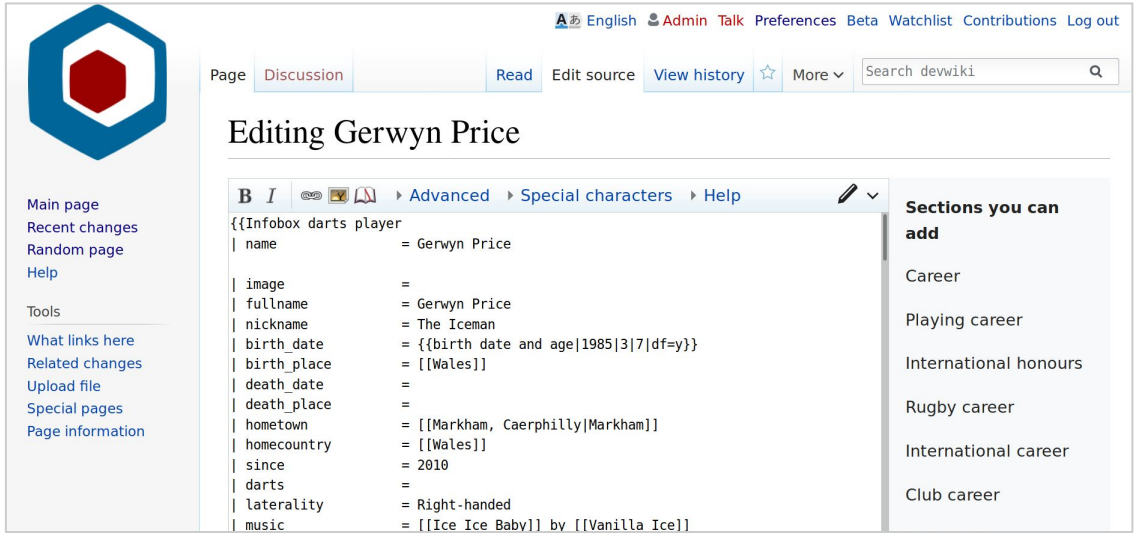


Surfacing recommendations

Created an API for retrieving category-based recommendations.


```
0:
  0: "Early life"
  1: 0.13062492806997353
1:
  0: "Career"
  1: 0.12751755092645875
2:
  0: "Personal life"
  1: 0.12314420531706756
3:
  0: "Early life and education"
  1: 0.12268385314765796
4:
  0: "Biography"
  1: 0.1115203130394752
```


Integrated section recommendations into MediaWiki via a Gadget.

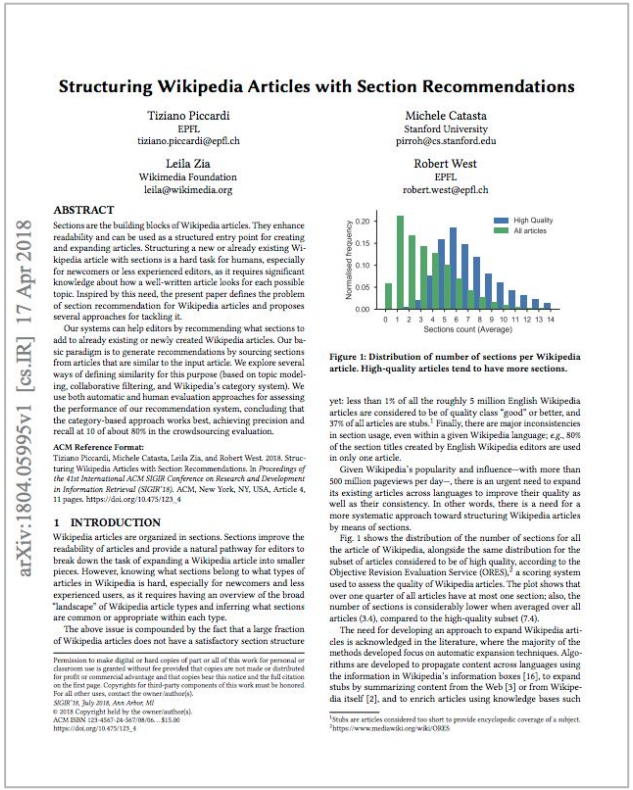


Other accomplishments

We presented preliminary results from this line of work at the [Wikimedia Research showcase]() in March.

We also [presented]() this research at WikiIndaba.

A [first paper]() resulting from this work has been accepted for publication at **SIGIR 2018** and will be presented at the conference in Ann Arbor, MI in July.



arXiv:1804.05995v1 [cs.IR] 17 Apr 2018

Structuring Wikipedia Articles with Section Recommendations

Tiziano Piccardi
EPFL
tiziano.piccardi@epfl.ch

Michele Catasta
Stanford University
pirroh@cs.stanford.edu

Leila Zia
Wikimedia Foundation
leila@wikimedia.org

Robert West
EPFL
robert.west@epfl.ch

ABSTRACT

Sections are the building blocks of Wikipedia articles. They enhance readability and can be used as a structured entry point for creating and expanding articles. Structuring a new or already existing Wikipedia article with sections is a hard task for humans, especially for newcomers or less experienced editors, as it requires significant knowledge about how a well-written article looks for each possible topic. Inspired by this need, the present paper defines the problem of section recommendation for Wikipedia articles and proposes several approaches for tackling it.

Our systems can help editors by recommending what sections to add to already existing or newly created Wikipedia articles. Our basic paradigm is to generate recommendations by sourcing sections from articles that are similar to the input article. We explore several ways of defining similarity for this purpose (based on topic modeling, collaborative filtering, and Wikipedia's category system). We use both automatic and human evaluation approaches for assessing the performance of our recommendation system, concluding that the category-based approach works best, achieving precision and recall at 10 of about 80% in the crowdsourcing evaluation.

ACM Reference Format:
Tiziano Piccardi, Michele Catasta, Leila Zia, and Robert West. 2018. Structuring Wikipedia Articles with Section Recommendations. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'18)*. ACM, New York, NY, USA, Article 4, 11 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Wikipedia articles are organized in sections. Sections improve the readability of articles and provide a natural pathway for editors to break down the task of expanding a Wikipedia article into smaller pieces. However, knowing what sections belong to what types of articles in Wikipedia is hard, especially for newcomers and less experienced users, as it requires having an overview of the broad "landscape" of Wikipedia article types and inferring what sections are common or appropriate within each type.

The above issue is compounded by the fact that a large fraction of Wikipedia articles does not have a satisfactory section structure yet: less than 1% of all the roughly 5 million English Wikipedia articles are considered to be of quality class "good" or better, and 37% of all articles are stubs.[1] Finally, there are major inconsistencies in section usage, even within a given Wikipedia language; *e.g.*, 80% of the section titles created by English Wikipedia editors are used in only one article.

Given Wikipedia's popularity and influence—with more than 500 million pageviews per day—, there is an urgent need to expand its existing articles across languages to improve their quality as well as their consistency. In other words, there is a need for a more systematic approach toward structuring Wikipedia articles by means of sections.

Fig. 1 shows the distribution of the number of sections for all the article of Wikipedia, alongside the same distribution for the subset of articles considered to be of high quality, according to the Objective Revision Evaluation Service (ORES),[2] a scoring system used to assess the quality of Wikipedia articles. The plot shows that over one quarter of all articles have at most one section; also, the number of sections is considerably lower when averaged over all articles (3.4), compared to the high-quality subset (7.4).

The need for developing an approach to expand Wikipedia articles is acknowledged in the literature, where the majority of the methods developed focus on automatic expansion techniques. Algorithms are developed to propagate content across languages using the information in Wikipedia's information boxes [16], to expand stubs by summarizing content from the Web [3] or from Wikipedia itself [2], and to enrich articles using knowledge bases such

SIGIR'18, July 2018, Ann Arbor, MI

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

Figure 1: Distribution of number of sections per Wikipedia article. High-quality articles tend to have more sections.

[1] Stubs are articles considered too short to provide encyclopedic coverage of a subject.
[2] https://www.mediawiki.org/wiki/ORES

In collaboration with: Tiziano Piccardi, Robert West (EPFL), Michele Catasta (Stanford)

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.

Identifying unsourced statements

1) We designed an annotation workflow to **label unsourced sentences in need of a citation** and implemented the interface in WikiLabels.

- Labeled data collection in English, French and Italian Wikipedia is currently underway (Q4)

In collaboration with: Besnik Fetahu (University of Hannover)

Identifying unsourced statements

2) We designed a **machine learning model to identify sentences in need of a citation**, based on multilingual natural language processing.

- Compiled a summary of *citation needed* rules
- Derived and implemented features reflecting these rules
- Created models using supervised machine learning
- Performed early tests with promising performance

Characterizing how readers use citations

We designed two research projects and a schema for instrumenting links and **characterizing how readers use citations**.

- Implementation, data collection and early analysis is planned in Q4

In collaboration with researchers at: Stanford University, EPFL, USU

Other accomplishments

We published **a 15-million record dataset of all publications with identifiers cited in Wikipedia articles in 300 languages**. We extracted data the top-cited books and scientific publications, as well as citation patterns over time.

The data has been already reused and analyzed by librarians and by organizations driving digitization and open access efforts. It was featured by *Wired* in May.

Other accomplishments

We submitted a WikiCite track proposal at the **Wikimedia Hackathon** in Barcelona and a workshop proposal at **Wikimania 2018** in Cape Town.

Both proposals got accepted.

We submitted a 3-year **funding proposal for the WikiCite series** (and satellite events) to the Alfred P. Sloan Foundation.

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 Sandbox · Wikimedia Foundation · Wikimedia Deutschland