

# **Analytics Quarterly Review**

Q4 2014

# AGENDA

---

**Introduction** (5 mins)

**Strategy** (30 mins)

**Development/Q&A** (35 mins)

**Break** (5 mins)

**Research & Data/Q&A** (35 mins)

**Conclusions & Asks/General Q&A** (10 mins)

# Asks of Audience

---

- Understand our evolving strategies to meeting the Foundation's Analytics needs
- Review the status of our Development initiatives
- Review our Research goals and achievements
- Appreciate where we could use some help

# Introduction

# Reflection

---

Q2 2014 Quarterly Review

"Everything I thought after 2 months was wrong"

Q3 2014 Quarterly Review

"I'm excited about what the team can accomplish (but there's a lot to do)"

Now

"We're making real progress and I'm excited about the direction"

# Group Structure

---

## **Development**

Builds the infrastructure, tools and datasets that enable the organization and the community to easily access, process and act on our data in a way that is consistent with our values.

## **Research and Data**

Supports the organization in making research-informed decisions, to better understand our editor community and projects, and to determine the impact of new programs and products that the Foundation is designing.

# Operating Model

---

- Research & Data is primary stakeholder contact and develops insights from our data
- Development builds out robust and scalable infrastructure
- R & D and Dev teams collaborate on many levels
  - Metric Standardization
  - Product Requirements
  - Privacy
- We are revisiting our R & D operating models to better serve the Foundation
  - More on this later in the presentation

# Follow up to Q3 Themes

---



[1]

## Impact

- Metrics standardized around Editor Model
- Research and Data expanding embedded and consulting models



[2]

## Focus and Balance

- Development scope streamlined with new Product Manager
- Research focusing activities on critical Foundation issues



[3]

## Community

- Hackathon connected us with Community members
- Research maintains high level of contact with external Researchers



## Collaboration

- First off-site created new levels of cohesiveness
- Standardized Metrics and Editor Engagement Vital Signs illustrate collaboration

[1] <https://www.flickr.com/photos/docman/3120465205/>

[2] <https://www.flickr.com/photos/quinnanya/5893328472/>

[3] <http://blog.wikimedia.org/2012/02/07/first-san-francisco-meetup-of-2012>



# Offsite Report

---

- The Analytics Team (Development and Research & Data sans Oliver) met outside Zurich for an off-site in May 2014
- This was the first time our distributed team had been together!
- We did some team building, development and worked on some key issues that impact all of us:
  - Team Values
  - Privacy
  - 1-3-5 Year Roadmaps



# Balancing Privacy and Understanding

---

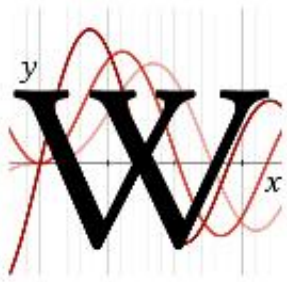
- Issues such as tracking, cookies and data collection are important to the Community and the Foundation
- Goal is to facilitate a balance between collecting the data we need to run a top 10 Internet site and respecting the wishes of privacy sensitive community members
- Unique Users and other important metrics depend on some aspects of this work
- The Analytics team has [summarized](#) our views and plans to work with Product on next steps next Quarter

## A bit more...

---

The model breaks combines the effects of disparate features into one easy to understand model that can be used by almost anyone to understand and drive Active Editors.

One of the key strengths of the model is that it provides a mechanism to connect insights around low-level metrics like edit counts into the high level metrics that measure success.



**Strategy**



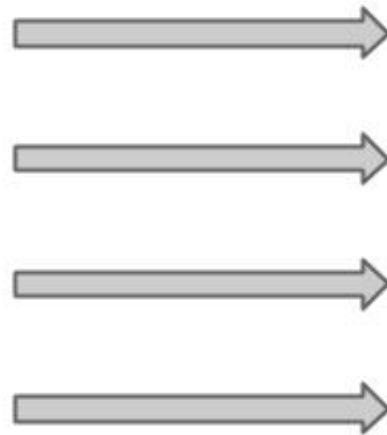
# Proposed Direction

---

- Provide the infrastructure, best practices and training around data at the Foundation
- Leverage our ongoing efforts
- Transition from pull to push and post to pre

# How do we stack up?

Classic DIKW Model



Analytics Offerings



[1] [http://en.wikipedia.org/wiki/DIKW\\_Pyramid#mediaviewer/File:DIKW-diagram.png](http://en.wikipedia.org/wiki/DIKW_Pyramid#mediaviewer/File:DIKW-diagram.png) (CC-BY-SA 3.0)

# Systems/Services View

---

## Infrastructure

- Production Slaves
- Server/Varnish Logs
- EventLogging
- Hadoop
- WebStatsCollector
- Various cron jobs and processing scripts

## Reporting and Visualization

- WikiStats
- WikiMetrics
- Limn

## Models and Techniques

- Editor Model
- Instrumentation Consultation
- Experimental Design
- Ad-hoc Research
- Predictive Modeling
- Training

## Research and Analysis

- Independent Research
- Literature Review
- Metric Design
- Analysis

# Meeting the Challenge

---

- Model/Techniques
  - Editor Model created and socialized
- Infrastructure
  - EventLogging Transitioned into team
  - Progress on Kafka/Hadoop
- Reporting/Visualization
  - Wikistats
  - WikiMetrics
  - Editor Engagement Vital Signs
- Research and Data
  - Operating Model tuning
  - Training
  - Analysis

**Deep Dive**

**Development**

**Research and Data**



# Editor Model

# A Common Understanding

---

**Challenge:** We need simple, actionable metrics for Editor behavior to understand our impact

**Solution:** A simple model of Editor activity that enables us to link features to our goals

# The Model

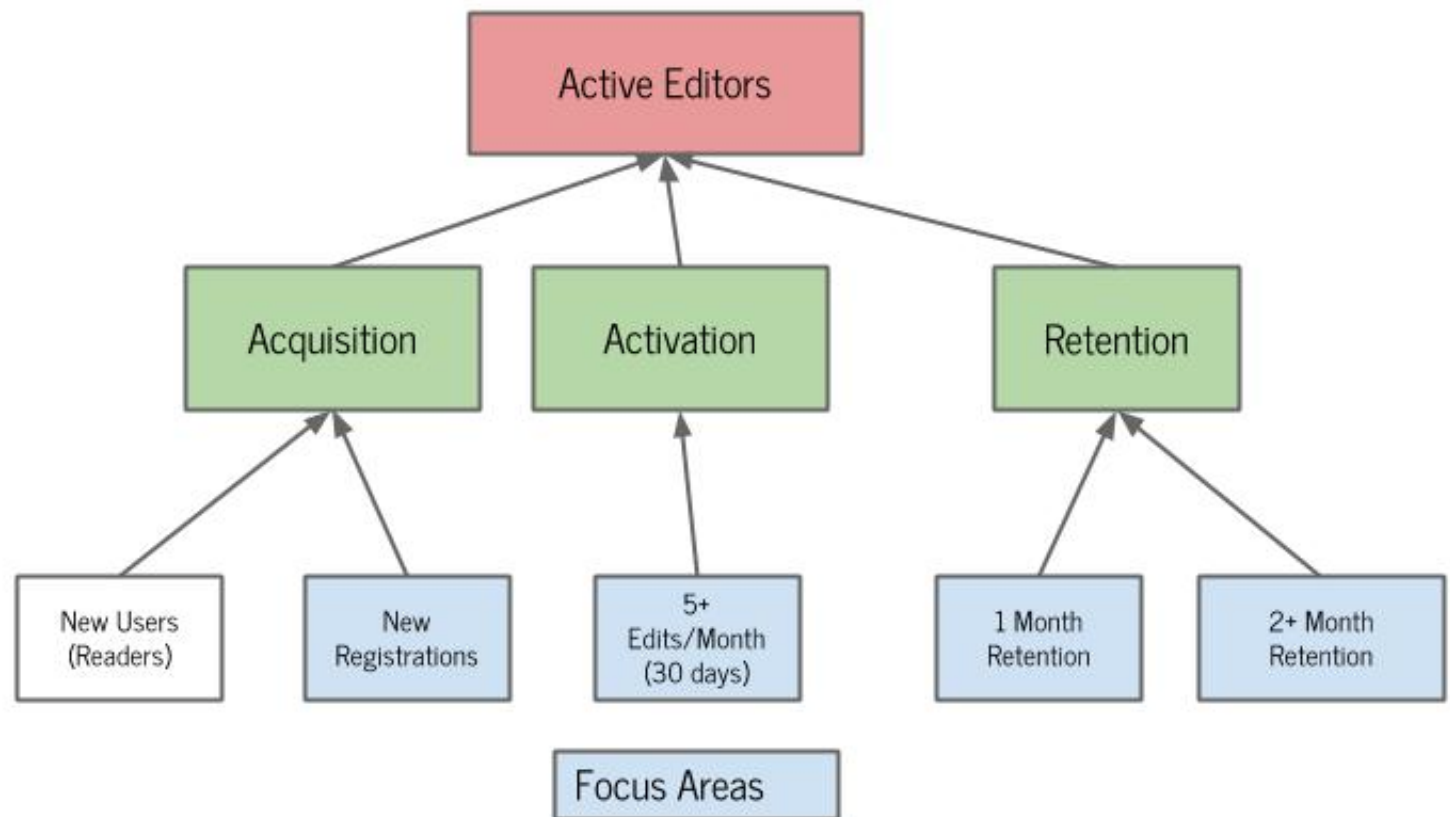
---

**Goal:** Increasing the number of Active Editors (5+ Edits/30 Days)

**Metrics:** Performance measurements that drive top level goal

**Levers:** Actionable metrics for Feature teams

There are an infinite # of levers – blue describes the focus areas



## A bit more...

---

The model breaks combines the effects of disparate features into one easy to understand model that can be used by almost anyone to understand and drive Active Editors.

One of the key strengths of the model is that it provides a mechanism to connect insights around low-level metrics like edit counts into the high level metrics that measure success.

# Progress

---

- Iterated with Product and User Research
- New Metrics defined by Research
  - Original Metrics still relevant and will be implemented in successive phases
- Metrics in MVP of Editor Engagement Vital Signs
- Used by Growth as Template for Goals
  - Will be used by Product Teams as goals develop in Q1 2015
- Socialized with Grantmaking
  - Valuable outside of Product and the Foundation

## Still to do

---

- Release Planning
- Visualization
- Prioritization of post-MVP Metrics
- Community Engagement



**Development**

# Q4 Narrative

---

## New Product Manager

focused on a few priorities  
productivity is becoming predictable

## We lost 2 team members

## We got a lot done!

Production Issues  
Editor Engagement Vital Signs  
EventLogging Transition (New)  
Kafka/Hadoop



# Product/Project Management Tooling

---

Abandoned Mingle 

Experimented with Phabricator 

Currently using ScrumBugs 

ScrumBugs to manage the backlog and sprints

Bugzilla for bugs

Etherpad to collaborate on tasking & estimating

Spreadsheet to track daily work

# Process Improvement

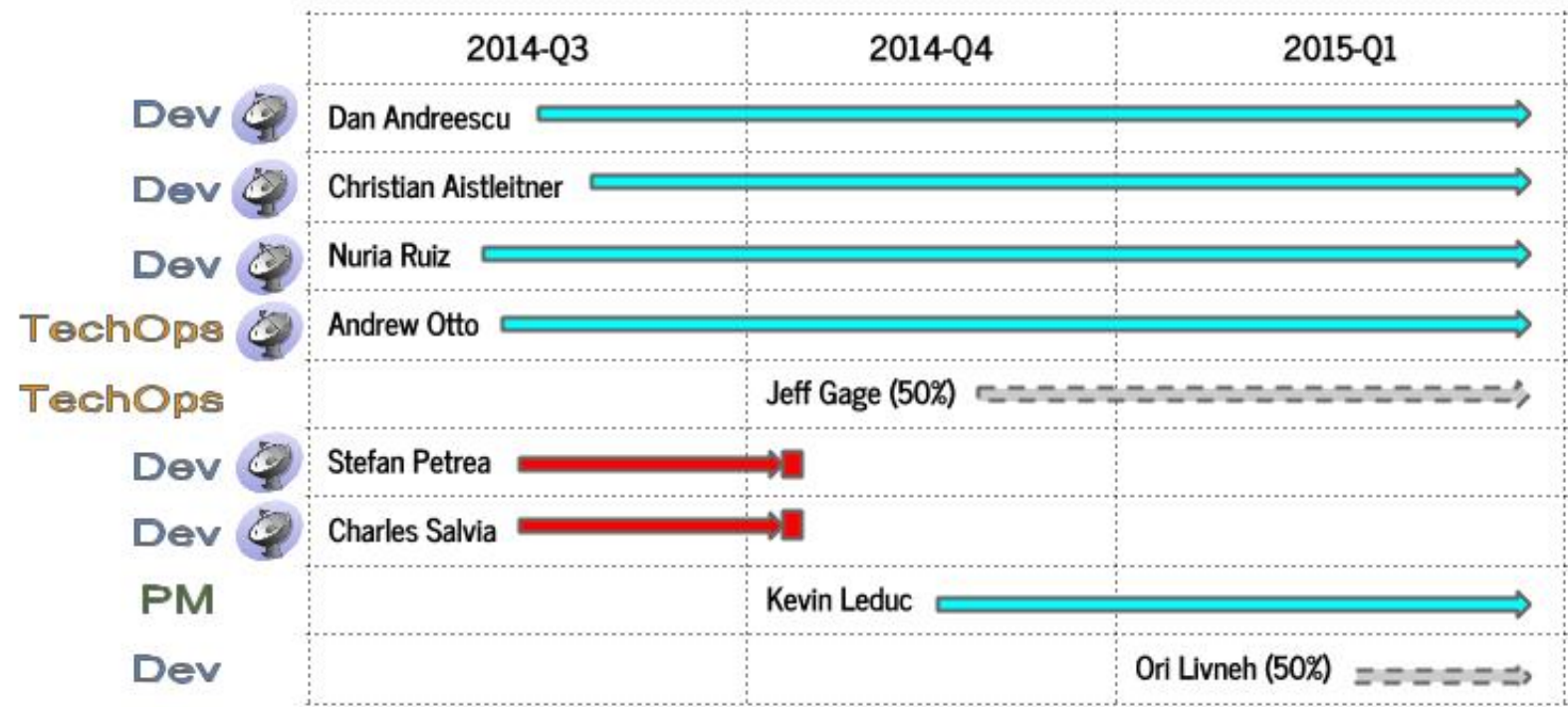
---

## Release Planning

Team's process for Sprint Planning is mature

We need to be able to answer when products will be released

# Staffing



# What we planned for Q4

---

Priority	Epic
0	Production Issues
1	Editor Engagement Vital Signs (EEVS)
2	Metrics about Mobile Usage
3	Dashboard Discoverability/Annotations
4	Event Logging Transition
5	Pageview API
6	Metric Definition Standardization (Dev/Page Views)
7	Accurate Pageviews for Wikipedia Zero
8	Simplify Limn Dashboard Deployment

# What We did in Q4

---

Priority	Epic
0	<b>Production Issues</b>
1	<b>Editor Engagement Vital Signs</b>
2	<b>Metrics about Mobile Usage</b>
3	Dashboard Discoverability/Annotations
4	<b>EventLogging Transition</b>
5	Pageview API
6	<b>Metric Definition Standardization (Dev/Page Views) (shared with Research)</b>
7	<b>Accurate Pageviews for Wikipedia Zero (support / bug fixes)</b>
8	Simplify Limn Dashboard Deployment
	<b>Refinery (Hadoop / Kafka)</b>

# Metrics on What We Did

---

	Completed Stories (Features)	Defects Fixed (Bugzilla Count)	Developers
EEVS / Wikimetrics	13	21	Dan, Nuria, Christian
Production Issues		18	Christian & team
EventLogging	1	2	Nuria, Christian, Dan
Refinery	4*	1*	Andrew, Christian

The major effort was on EEVS

Significant amount of time spent on production issues (unplanned work)

EventLogging is ramping up following its transition to our team

\* Refinery is in development and we are beginning to formalize tracking

# Production Issues

---

Unplanned support required for infrastructure or dashboards.

Database  
Slaves

Wikipedia  
Zero

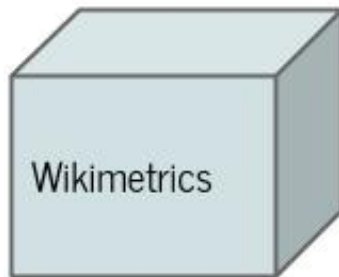
Geowiki

Dashboards &  
Report Cards

stats.grok.se

# Editor Engagement Vital Signs

---



## Wikimetrics:

User friendly platform for generating reports (data)

Originally created for Grantmaking to measure program success

Outputs Standardized Metrics (defined by Research)



# Editor Engagement Vital Signs

---

## Enhancements to **Wikimetrics** platform

- Recurring reports

- Public reports

- Running a report on an entire project (e.g. enwiki or dewiki)

## Editor Model Metrics

- Newly Registered User

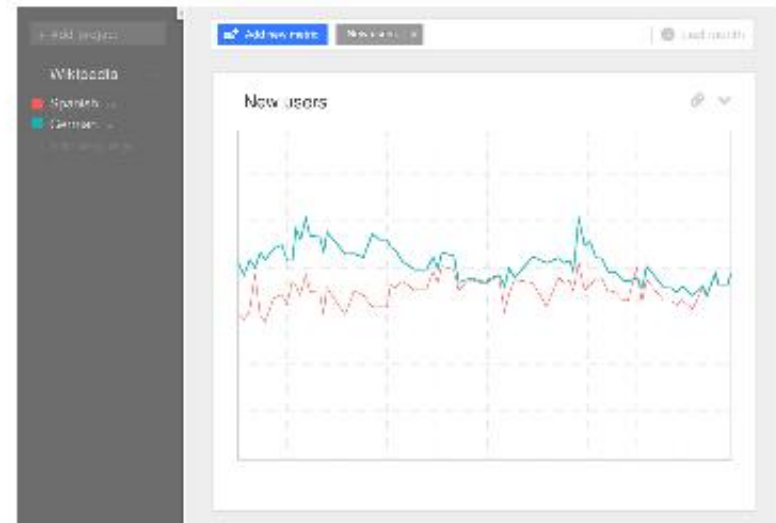
- 3 other metrics are ready to begin implementation

# Editor Engagement Vital Signs

---

## Dashboards

UX wireframes and mockups



We expected to implement components of a dashboard this quarter and did not have the bandwidth to do so.

# Metrics about Mobile Usage

---

Work shared with the research team

Oliver Keyes (Research) did the lion's share

Dev team supported with

- Selecting a User Agent Parser

- Using Hadoop (It's still brittle)

# EventLogging Transition

---

Officially took over support of EventLogging from the Platform Team

Implemented EventLogging monitoring  
monitors the throughput of events

Discussions / Deep Dives

Privacy

Data retention (aggregation, pruning) & purging

Sampling Events

Developing the backlog

scope: operationalization (not new features)

# Metric Definition Standardization

---

Shared with Research

Implementation is part of EEVS

# Accurate PageViews for Wikipedia Zero

---

Support & Bugfixes

treated as "Operational Issues"

Wikipedia Zero's team has updated their infrastructure

Requires updates on our end

Work has been tasked out

Dependencies on Hadoop & ETL

# Refinery (Hadoop/Kafka) Work

---

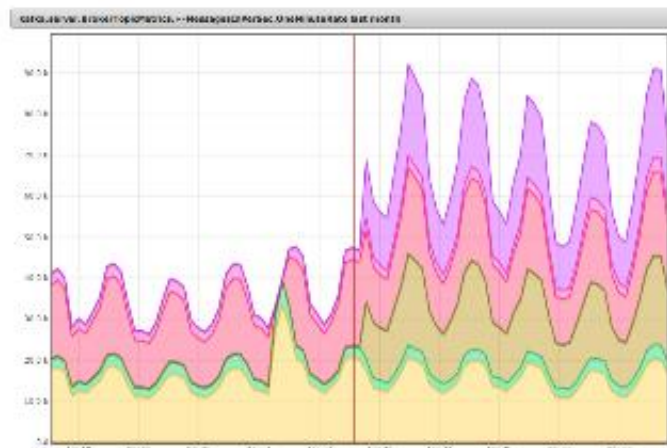
Collaboration of Tech-ops & Development on key infrastructure

Ingesting Text, Mobile, Images and Bits request logs

➔ 150K/RPS through Kafka into Hadoop

Camus productized

Capacity Planning/Hardware support



Turning on Text ingestion

## Projects Supported

Mobile Metrics

Wikipedia Zero

PageView API (depends on Research's PageView Definition)

# Other Things We Worked On

---

## Grantmaking

Enhancements to Wikimetrics

## Offsite / Hackathon

## Data-center migration

## Hiring

## Geowiki

New hiring req for work on dashboards



# Q1-2015 goals

Editor Engagement Vital Signs

EventLogging

Refinery

Support

# Data Map *(Draft)*

Category	Target Audience	Example	Current Implementation	Target Implementation	Migration complexity
Operating Model	Execs, Press	Active Editor model	Wikimetrics Prod DBs	Wikimetrics	Medium
Traffic	Execs, Press, Community, Product	Page Views, Unique Users, Breakdowns	WikiStats Comscore	Hadoop	High
Fundraising	Execs	\$, # Donors, % donating	FR Tech	Data from Hadoop	Low (when Hadoop is operational)
Project Level Metrics	Product Community Grantmaking	Newly Registered Users New Pages # of Edits	EventLogging	EventLogging, Dashboard Organization, Additional Instrumentation	Medium
Feature Metrics	Product MediaWiki Core	Clicks, Funnel Performance	EventLogging	EventLogging, Dashboard Organization, Additional Instrumentation	High

# Q1-2015 Prioritization

---

**We are still reviewing yearly goals of our stakeholders and have not prioritized the following projects**

**Dashboard Discoverability / Annotations**

**Wikipedia Zero**

**Geowiki**

# Q1-2015 Goals

---

## Editor Engagement Vital Signs

Complete the MVP (dashboard & metrics related to the high level editor lifecycle model)

Dashboard Technology Review

## Event Logging

Operationalization (prune/scrub/purge and aggregate some of the data, complying with privacy policy & data retention guidelines)

geocoding IPs

# Q1-2015 Goals

---

## Refinery

- CDH 5 (New Hadoop Release)

- Capacity Enhancement

- Kafka Hardening

- ETL: UA Processing

## Wikistats

- Enhancements and bug-fixes

## Support

- Support existing legacy and current systems in production

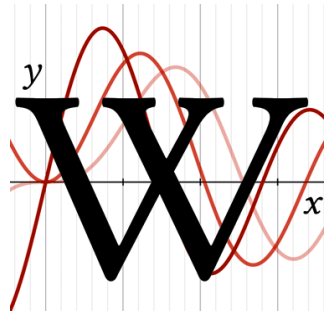
- Continue to scale team



**Questions?**



Break



# Research and Data



# Outline

who we are / what we do

Q4-2014 retrospective

Q1-2015 goals

H1-2015 staffing outlook

# Who we are



Erik Zachte



Aaron Halfaker



Oliver Keyes



Dario Taraborelli



Leila Zia

2013

2014

2015



Q4



Q1



Q2



Q3



Q4



Q1

# What we do

We apply a range of research methods to produce knowledge on our users and our projects and support decision-making, product evaluation and strategy at the Foundation and within the movement.

DATA MINING

BEHAVIORAL ANALYSIS

DATA MODELING

PREDICTIVE MODELING

CONTROLLED EXPERIMENTATION, A/B TESTING

EXPLORATORY RESEARCH

RESEARCH CONSULTING

# Q4-2014 retrospective

Metrics standardization

Topical research

Team support / community outreach

# Q4 goals

A. Deliver stage-2 metrics

B. Topical research

C. Support focus areas (growth, mobile, fundraising)

D. Ad-hoc consulting for other teams

# Metrics standardization

# Metrics Standardization: Editor Model

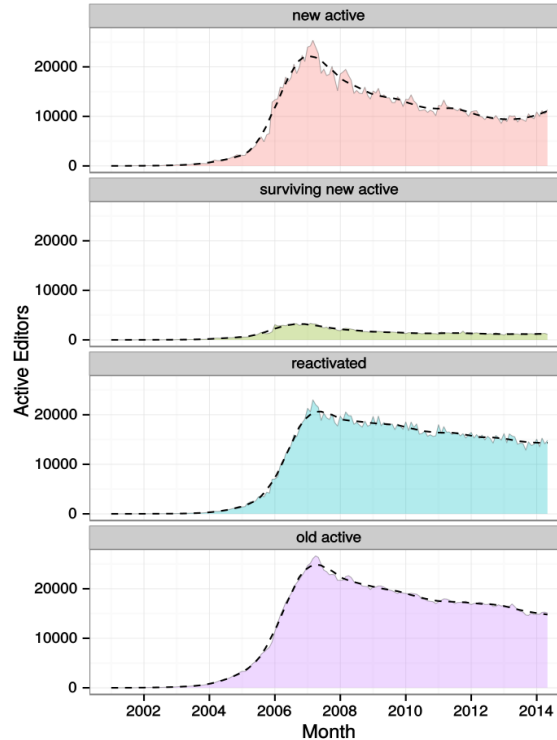
activation

short-term retention

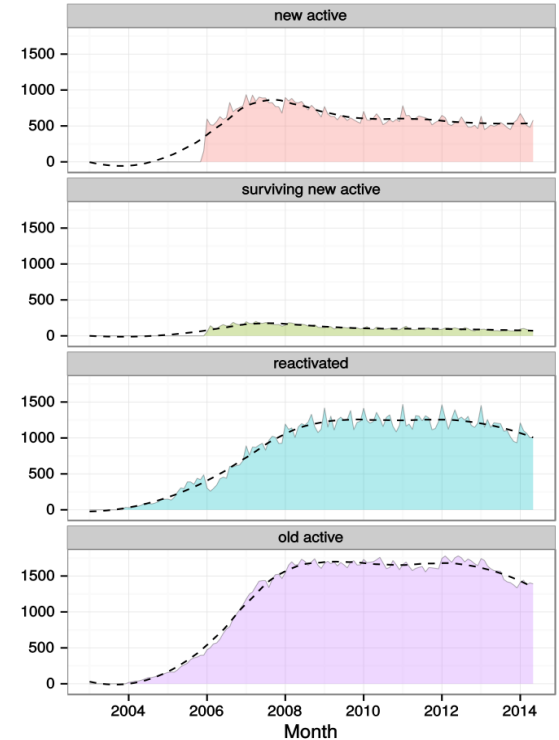
reactivation

long-term retention

English Wikipedia



Italian Wikipedia




Monthly Active Editors

[https://meta.wikimedia.org/wiki/Research:Rolling\\_monthly\\_active\\_editor](https://meta.wikimedia.org/wiki/Research:Rolling_monthly_active_editor)

# Metrics Standardization: Editor Model

Research:Rolling monthly active editor

 **This page in a nutshell:** This page describes an active editor metric computed on a daily basis using a 30-day rolling window.

**Contents** [\[hide\]](#)

- 1 About
- 2 Acquisition and short-term retention
  - 2.1 New active editors
  - 2.2 Surviving new active editors
- 3 Long-term retention and reactivation
  - 3.1 Recurring old active editors
  - 3.2 Reactivated editors
- 4 Assumptions
- 5 Notes

**The Editor Model™**  
one metric to rule them all  
Toby Night - Aaron Halflaker - Dario Taraborelli  
June 17, 2014

A high-level overview of the design of Rolling Monthly Active Editors

**About** [\[edit\]](#)

The Wikimedia Foundation has been using [monthly active editors](#) as the key metric to track the size of the **registered user population of contributors** to Wikimedia projects. This page documents a new metric tracking monthly active editors and:

- computed using a rolling 30-day window instead of calendar months
- broken down into different segments of the registered user population.

Generally speaking, in any given month, three categories of users are included in the active editor definition:

- **New users** who **became active** for the first time in the current month;
- Recently registered users who **remained active** (or survived) in the current month;
- Previously registered users who either **remained active** or were **reactivated** in the current month.

These informal categories can be operationalized as follows:

**Acquisition and short-term retention** [\[edit\]](#)

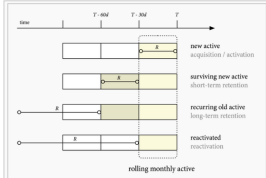
The first two categories measure the acquisition and short-term retention of active editors.

**New active editors** [\[edit\]](#)

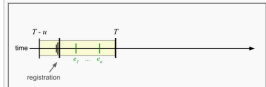
*Main article: R:Rolling new active editor*

A rolling new active editor( $T, u, n$ ) is a **newly registered user** who both registered and completed  $n$  **edits to pages** in any **namespace** of a Wikimedia project between  $T - u$  and  $T$ .

- $n = 5$  edits
- $u = 30$  days



A diagram illustrating the composition of the rolling monthly active editor metric.



**New active editor (rolling)**

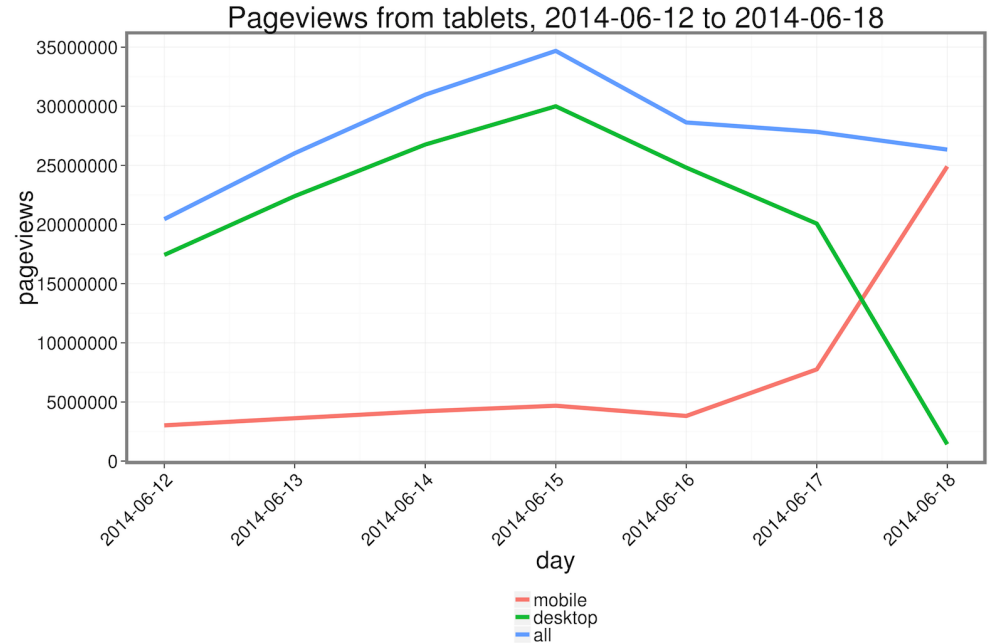
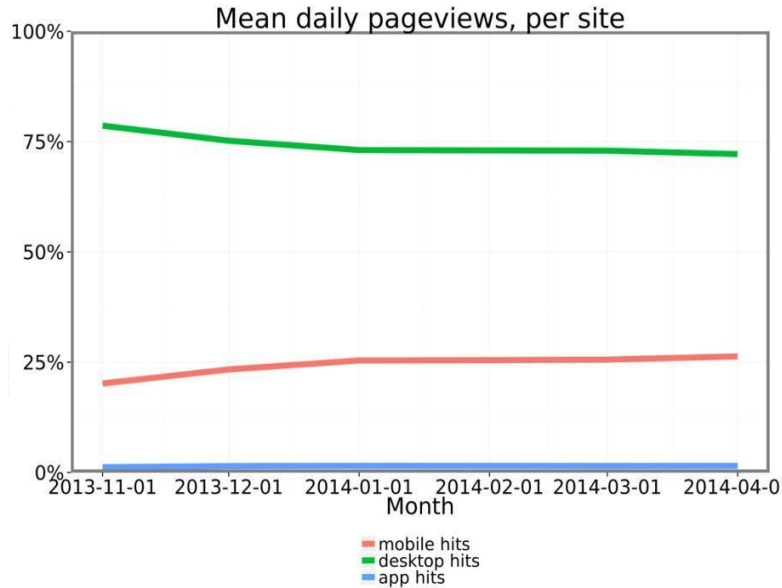
[https://commons.wikimedia.org/wiki/File:Editor\\_Model\\_review.pdf](https://commons.wikimedia.org/wiki/File:Editor_Model_review.pdf)



Topical research

# Topical research: Mobile reach

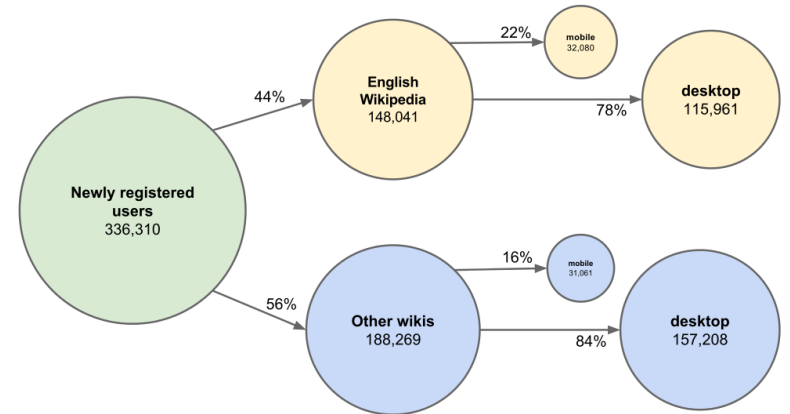
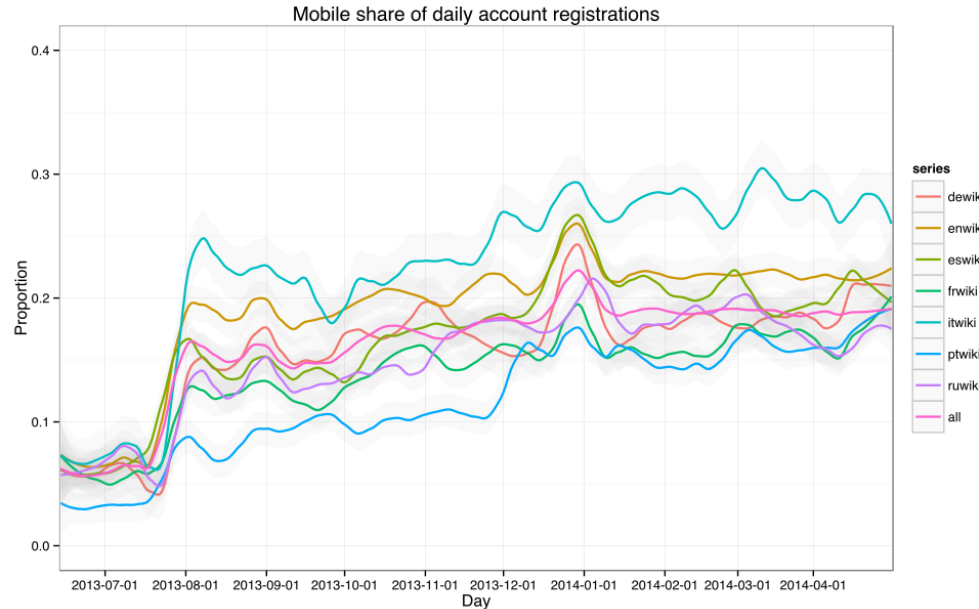
## How many visitors can we reach on mobile?



[https://www.mediawiki.org/wiki/File:2014-05-01\\_Mobile\\_Metrics.pdf](https://www.mediawiki.org/wiki/File:2014-05-01_Mobile_Metrics.pdf)  
[https://meta.wikimedia.org/wiki/Research:Mobile\\_Traffic](https://meta.wikimedia.org/wiki/Research:Mobile_Traffic)

# Topical research: Mobile acquisition

## How many users can we acquire on mobile compared to desktop?



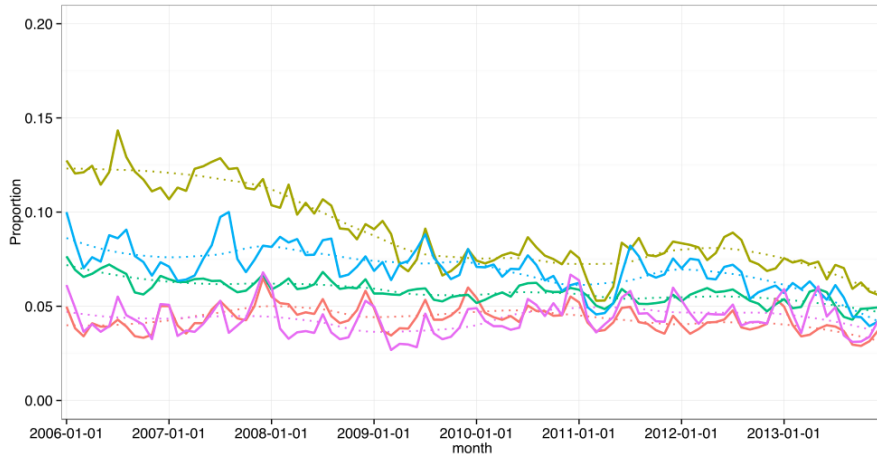
Data from Feb. 11 - Mar. 13 (30 days)

# Topical research: Editor activation

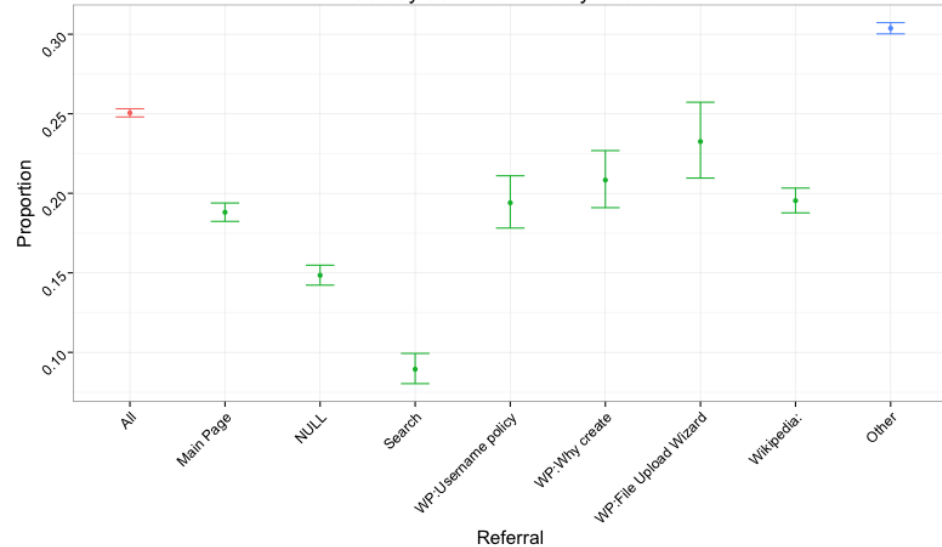
## Where do new (active) editors come from?

editor activation rate (5 edits in 24h, all ns)

project — es — de — en — fr — pt

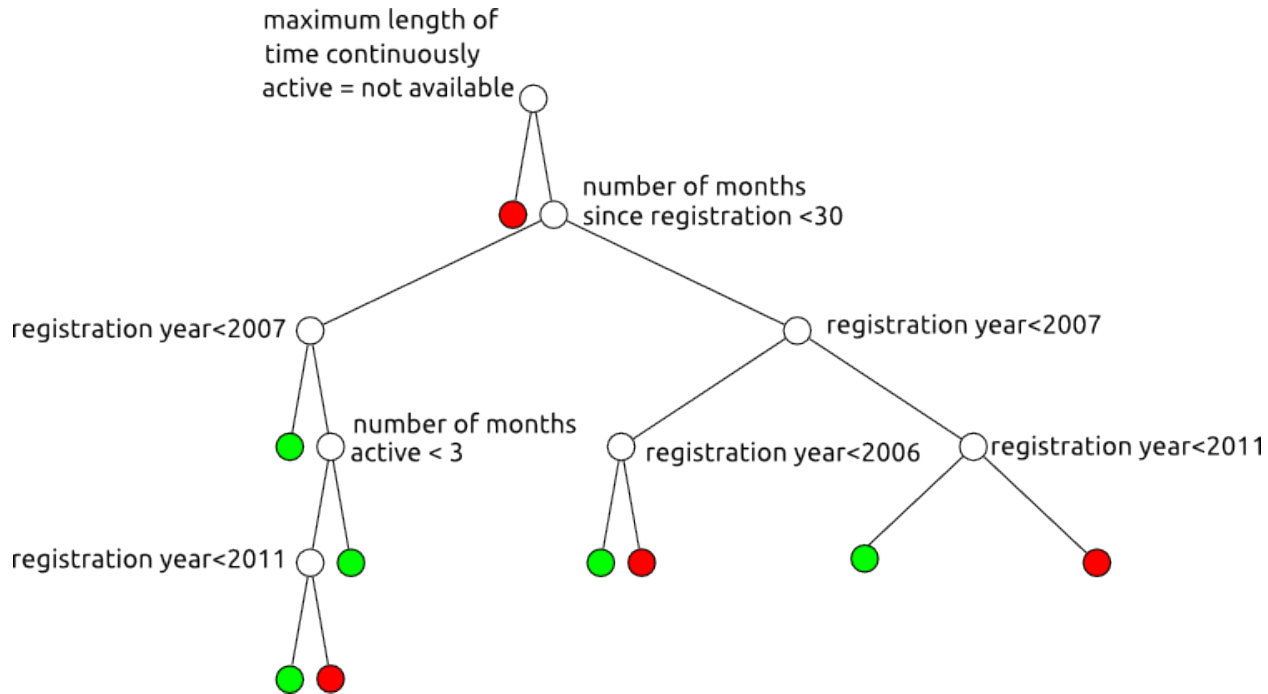


30 day activation rate by referral



# Topical research: Active editor retention / trajectories

What are the top predictors of short or long-term active editor survival?



# Topical research: Active editor migration

Do active editors *leave* or *migrate* to other projects?

```
This report is about a subset of editors who contribute to
only those with at least 50 edits in a year and at least 5

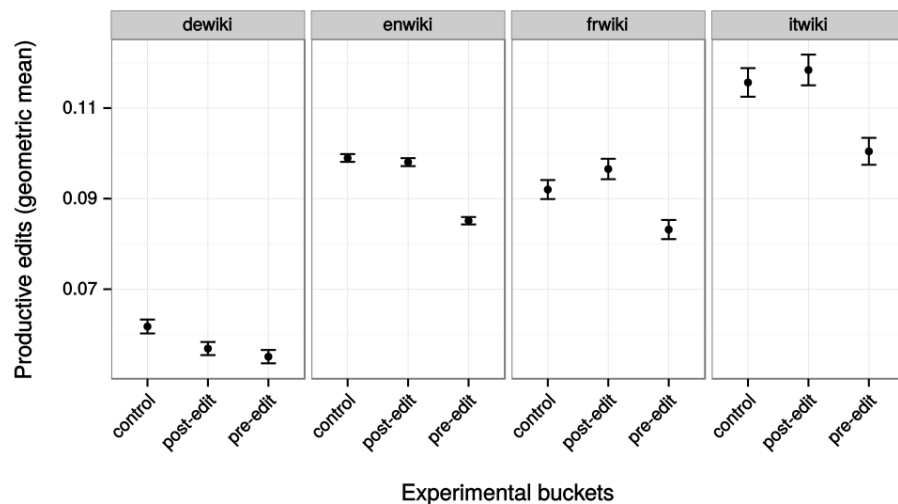
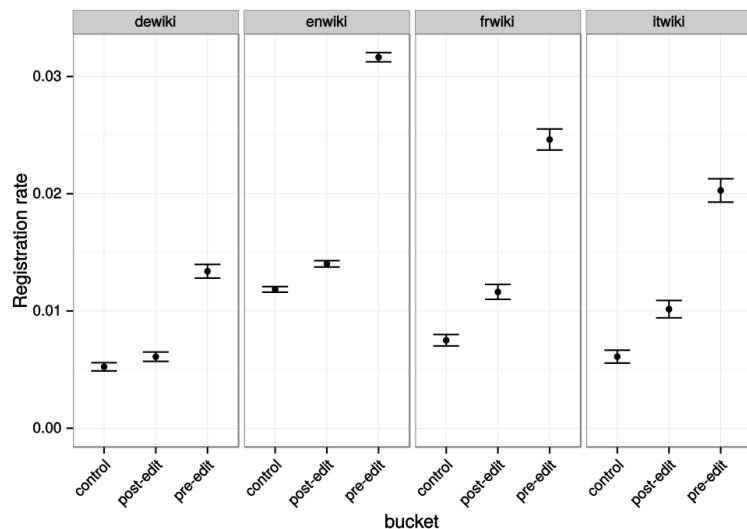
Migrations per path (= from project -> to project)
,wb:wikibooks, wk:wiktionary, wn:wikinews, wo:wikivoyage,

total      wp->wk   wk->wp   wp->wd   wp->ws   wp->wx   wp->wb   ws
2002              786     588     509     461     373     356
2003              3              6              1
2004              6              6              2              5              3
2005              32             26             14             10             31
2006              67             53             42             13             51
2007             100             64             50             32             63
2008              94             58             65             25             57
2009             130             63             49             34             41
2010              84             80             51             31             37
2011              74             76             48             49             21
2012              86             69              53             51             50             20
2013              74             60             359             59             58             19
```

<https://trello.com/c/3ecjp9aM/237-master-monthly-editor-activity-data>

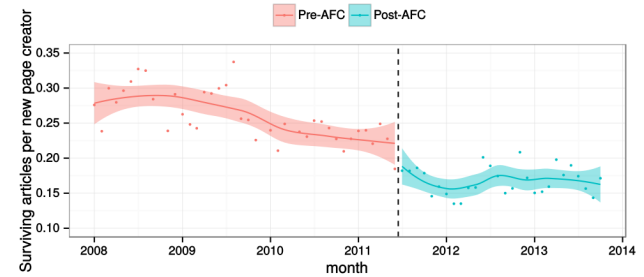
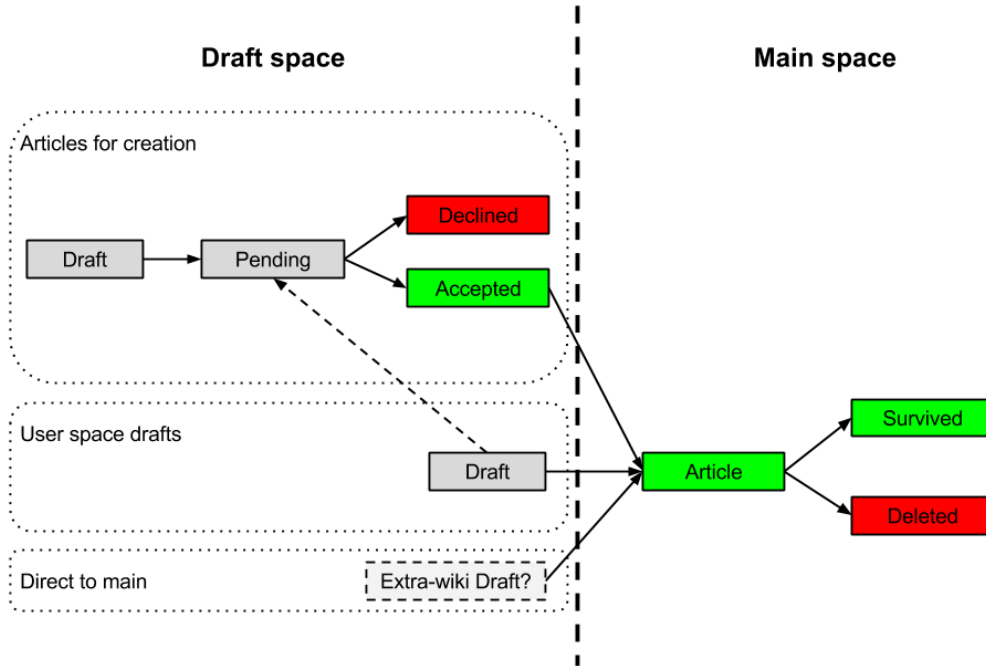
# Topical research: Anonymous editor acquisition

## How effectively can we acquire anonymous editors?



# Topical research: Article survival

## How article creation workflows (drafts, AfC) impact content growth





Team support, ad-hoc consulting  
and community outreach

# Teams supported: Focus areas

Q1

Growth

VE

Q2

Growth

Mobile

Q3

Growth

Mobile

Q4

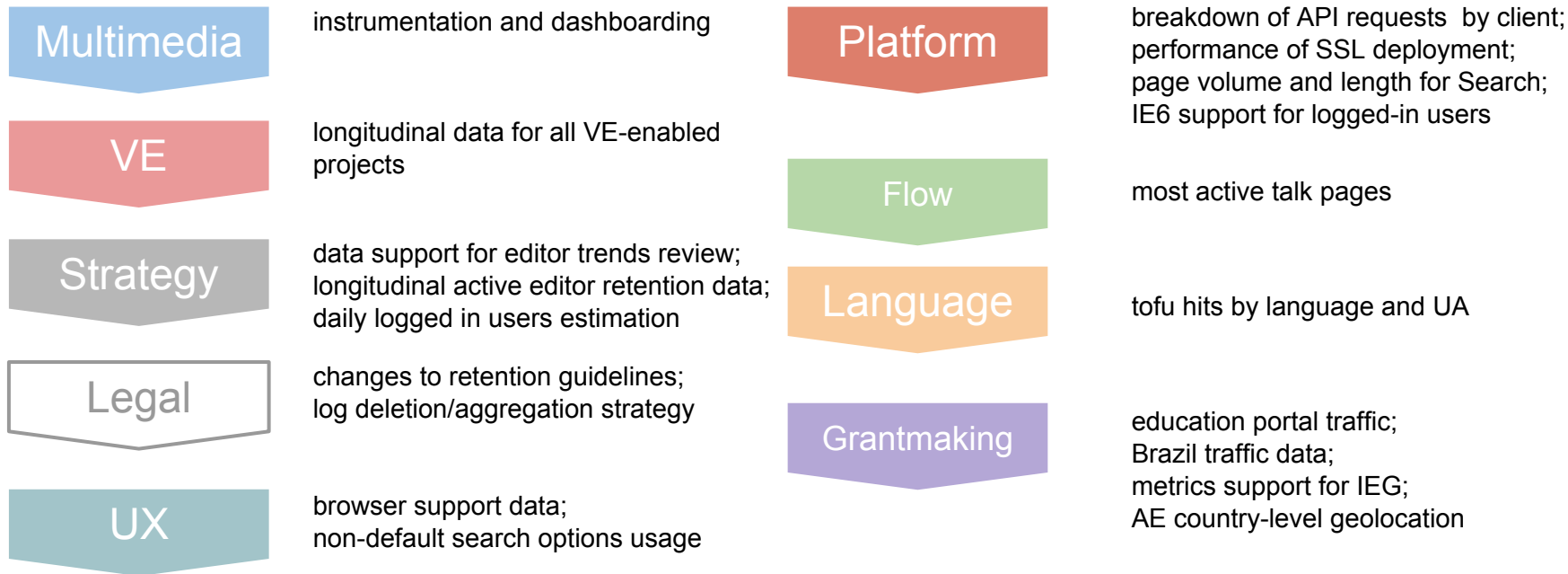
Growth

Mobile

Zero

Fundraising

# Teams supported: Consulting and ad-hoc analysis



# Community support

[Editor trends in Portuguese Wikipedia](#)

[Portuguese Wikibooks edit dashboards](#)

GLAM analytics ([NARA pilot](#))

Wikistats Portal overhaul

Various community requests @ Zurich hackathon

# Outreach

[WMF Research Showcase](#) (3 showcases, 6 presentations) + reading groups

2 conference papers submitted

1 accepted at WikiSym '15 (AfC process & productivity)

1 submitted to CSCW '15 (MoodBar)

[8 proposals](#) submitted and accepted at Wikimania '15

5 driven by Research & Data

3 co-authored

Mozilla UX / Research summit; Mozilla Science Labs talk

# Toolkits and documentation

[wikipediR](#)

R wrapper for the MediaWiki API

[WMUtils](#)

Utilities for geolookup and UA parsing of request logs

[mediawiki-utilities](#)

General data processing utilities in python

[mwoauth](#)

MediaWiki OAuth support for python tools

[wikiclass](#)

Automated quality assessment of Wikipedia articles

Analytics infrastructure documentation on Wikitech

([data access](#) - [geolocation](#) - [hive queries](#))

# Q1-2015 goals

R&D process

Metrics standardization

Topical research

Team support / community outreach

# Process

Uplevel team process (prioritization sprints, project management)

Finetune operating model (consulting vs embedded)

Push on horizontal integration with other research teams (UX, grantmaking)



# **Metrics standardization: Organizational alignment**

**Socialize the Editor Model and Vital Signs metrics**

**Consistent target setting across teams**

**Provide metrics training**

# Metrics standardization: Definitions and analysis



# Topical research, continued

Anonymous acquisition research

Mobile apps adoption/mobile activation trends

Predictive models of editor activation/retention

Unique visitors analysis; Readership metric definitions; mobile traffic trends

Cross-wiki migration

# Formal collaborations

Knowledge graph and recommender systems  
(GroupLens, UMN)

Traffic data anonymization and aggregation  
(Los Alamos National Laboratory)

# Teams supported: Focus areas

Q1-2015

Growth

Mobile

Fundraising

# H1-2015

## Staffing outlook

# Staffing



Erik Zachte



Aaron Halfaker



Oliver Keyes

Req 2  
(traffic)



Dario Taraborelli



Leila Zia

Req 1  
(FR)

2014

2015



Q1



Q2



Q3



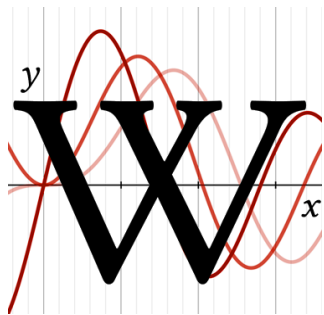
Q4



Q1

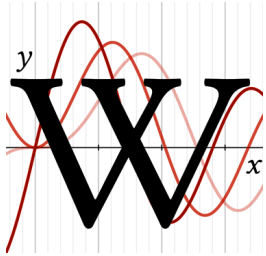


Q2

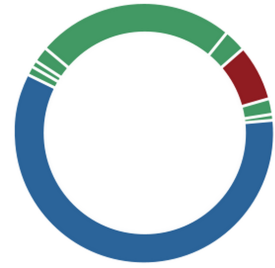


Questions?





# Conclusions



# Challenges

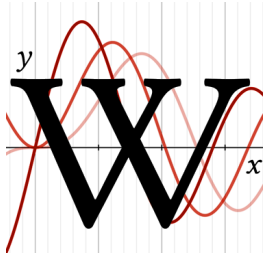
---

- **Staying Focused and Delivering on our Commitments**
- **Lower level tasks and legacy support sap resources**
- **Community Engagement**
- **Development Transparency**

# Asks

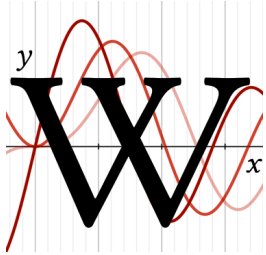
---

- 1. Project Manager/Scrum Master**
- 2. Tech Ops Support**
- 3. Exec Support for Metrics Standardization**
- 4. Operational Research (Interns)**



Questions?





Thank You

