Publishing Wikipedia project usage data with strong privacy protections and without tracking

Hal Triedman
23 June 2022

Table of contents

01 Introduction + context

The Wikimedia Foundation (WMF)

- Nonprofit
- Develops open-source software and hosts projects like Wikipedia, Commons, MediaWiki, Wikidata, etc.
- 22B pageviews per month
- 803 active projects, 316 languages, visitors in every country in the world
- Wikipedia is 7th-most visited site

WMF’s Open Access Policy

Differential privacy: Revision history

Help

View logs for this page (view filter log)

Filter revisions

External tools: Find addition/removal (Alternate) · Find edits by user (Alternate) · Page statistics · Pageviews · Fix dead links

For any version listed below, click on its date to view it. For more help, see Help:Page history and Help:Edit summary. (cur) = difference from current version, (prev) = difference from preceding version,
m = minor edit, → = section edit, ← = automatic edit summary
(newest | oldest) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)

Compare selected revisions

- (cur | prev) ◉ 11:18, 25 April 2022 HaeB (talk | contribs) . . (36,350 bytes) (+253) . . *(Restored revision 1080669717 by 192.184.205.102 (talk): Fully revert vandalism)* (undo) *(Tags: Twinkle, Undo)*
- (cur | prev) ◉ 07:25, 25 April 2022 Haskle (talk | contribs) . . (36,097 bytes) (+36) . . *(The first sentence was incomplete, probably something was deleted accidentally by previous editors.)* (undo) *(Tag: Reverted)*
- (cur | prev) ○ 06:13, 25 April 2022 37.39.165.173 (talk) . . (36,061 bytes) (−289) . . (undo) *(Tags: Reverted, Visual edit, Mobile edit, Mobile web edit)*
- (cur | prev) ○ 17:33, 2 April 2022 192.184.205.102 (talk) . . (36,350 bytes) (+4) . . *(→Group privacy: Small edit for improved clarity.)* (undo)
- (cur | prev) ○ 16:16, 10 March 2022 GhostInTheMachine (talk | contribs) . . (36,346 bytes) (−17) . . *(Changing short description from "System for publicly sharing information about a dataset" to "Methods of safely sharing general data" (Shortdesc helper))* (undo)
- (cur | prev) ○ 10:13, 10 March 2022 GhostInTheMachine (talk | contribs) . . (36,363 bytes) (−51) . . *(Undid revision 1076207980 by Dr. José Ivon Rodrigues da Cruz, Justiça Federal (talk))* (undo) *(Tag: Undo)*
- (cur | prev) ○ 03:00, 10 March 2022 Pan-m72 (talk | contribs) . . (36,414 bytes) (+32) . . *(→See also: Add link for Local differential privacy)* (undo)
- (cur | prev) ○ 22:52, 9 March 2022 Dr. José Ivon Rodrigues da Cruz, Justiça Federal (talk | contribs) . . (36,382 bytes) (+51) . . (undo) *(Tags: Reverted, Mobile edit, Mobile app edit, Android app edit)*
- (cur | prev) ○ 11:38, 8 March 2022 2a02:8108:973f:e934:3c68:7e74:5b3b:24fa (talk) . . (36,331 bytes) (0) . . *(→References)* (undo)
- (cur | prev) ○ 18:19, 3 March 2022 Gilliam (talk | contribs) **m** . . (36,331 bytes) (−27) . . *(Reverted edits by 2405:9800:BC12:55D8:C53E:6278:74D9:6876 (talk) to last version by Nicolas09F9)* (undo) *(Tag: Rollback)*
- (cur | prev) ○ 18:18, 3 March 2022 2405:9800:bc12:55d8:c53e:6278:74d9:6876 (talk) . . (36,358 bytes) (+27) . . *(→History)* (undo) *(Tags: Reverted, Visual edit)*
- (cur | prev) ○ 01:51, 22 December 2021 Nicolas09F9 (talk | contribs) . . (36,331 bytes) (+21) . . *(Undid revision 1060796822 by 78.39.19.15 (talk), vandalism)* (undo) *(Tag: Undo)*
- (cur | prev) ○ 18:40, 17 December 2021 78.39.19.15 (talk) . . (36,310 bytes) (−21) . . *(Delete clear all settings)* (undo) *(Tags: Reverted, Visual edit, Mobile edit, Mobile web edit)*
- (cur | prev) ○ 14:23, 17 December 2021 GatienC (talk | contribs) **m** . . (36,331 bytes) (+60) . . *(changed the ϵ to ε in math formulae for consistency)* (undo)
- (cur | prev) ○ 04:03, 17 December 2021 I dream of horses (talk | contribs) **m** . . (36,271 bytes) (+7) . . *(Random page patrol with AutoWikiBrowser, typo(s) fixed: et. al. → et al., 202-210 → 202–210 (6))* (undo) *(Tag: AWB)*
- (cur | prev) ○ 19:55, 15 December 2021 2601:481:8601:86c0:90fa:c968:ecc4:7646 (talk) . . (36,264 bytes) (−117) . . (undo) *(Tags: Visual edit, Mobile edit, Mobile web edit)*
- (cur | prev) ○ 20:58, 25 November 2021 Sjö (talk | contribs) **m** . . (36,381 bytes) **(+1,714)** . . *(Reverted edits by Hieulamq11 (talk) to last version by ClueBot NG)* (undo) *(Tag: Rollback)*
- (cur | prev) ○ 20:58, 25 November 2021 Hieulamq11 (talk | contribs) . . (34,667 bytes) **(−1,714)** . . *(Https//Microsoft.email.com@Gmail.com)* (undo) *(Tags: Reverted, Mobile edit, Mobile web edit, adding email address)*
- (cur | prev) ○ 05:28, 16 November 2021 ClueBot NG (talk | contribs) **m** . . (36,381 bytes) (−7) . . *(Reverting possible vandalism by 186.193.34.81 to version by Ffffrr. Report False Positive? Thanks, ClueBot NG. (4077870) (Bot))* (undo) *(Tag: Rollback)*
- (cur | prev) ○ 05:27, 16 November 2021 186.193.34.81 (talk) . . (36,388 bytes) (+7) . . *(Mestrao)* (undo) *(Tags: Reverted, Visual edit, Mobile edit, Mobile web edit)*
- (cur | prev) ○ 20:04, 3 November 2021 Ffffrr (talk | contribs) **m** . . (36,381 bytes) (+78) . . (undo) *(Tags: Mobile edit, Mobile app edit, iOS app edit)*
- (cur | prev) ○ 02:47, 30 October 2021 2a00:1fa2:82da:f82f:29da:5c28:735d:1396 (talk) . . (36,303 bytes) (+33) . . *(antonnsk23102014@gmail.com)* (undo) *(Tags: Visual edit, Mobile edit, Mobile web edit)*

Wikipedia
The Free Encyclopedia

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

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Atom
Special pages
Page information
Wikidata item

Languages

Monthly overview

All wikis

Reading

Total page views

22B

April ↓ *-4.82% month over month*

262B ↓ *-6.55% year over year*

Last 12 Months (May 2021 - Apr 2022)

Page views by country

Countries with the most views for April

3B	United States of A
987M	Japan
885M	Germany
796M	United Kingdom

The Unique devices metric is not available for all projects. Select a specific wiki

Contributing

Edits

46M

April ↓ *-11.31% month over month*

520M ↓ *-6.34% year over year*

Last 12 Months (May 2021 - Apr 2022)

New registered users

237K

April ↓ *-11.02% month over month*

3M ↓ *-18.74% year over year*

Last 12 Months (May 2021 - Apr 2022)

User edits

26M

April ↓ *-9.31% month over month*

287M ↓ *-2.08% year over year*

Last 12 Months (May 2021 - Apr 2022)

Content

Total media requests

71B

April ↓ *-8.59% month over month*

882B ↓ *-14.75% year over year*

Last 12 Months (May 2021 - Apr 2022)

Net bytes difference

20GB

April ↓ *-68.35% month over month*

409GB ↓ *-51.27% year over year*

Last 12 Months (May 2021 - Apr 2022)

Absolute bytes diff

25GB

April ↓ *-68.42% month over month*

491GB ↓ *-49.27% year over year*

Last 12 Months (May 2021 - Apr 2022)

Wikimedia REST API 1.0.0 OAS3

/api/rest_v1/?spec

This API provides cacheable and straightforward access to Wikimedia content and data, in machine-readable formats.

Global Rules

- Limit your clients to no more than 200 requests/s to this API. Each API endpoint's documentation may detail more specific usage limits.
- Set a unique `User-Agent` or `Api-User-Agent` header that allows us to contact you quickly. Email addresses or URLs of contact pages work well.

By using this API, you agree to Wikimedia's Terms of Use and Privacy Policy. Unless otherwise specified in the endpoint documentation below, content accessed via this API is licensed under the CC-BY-SA 3.0 and GFDL licenses, and you irrevocably agree to release modifications or additions made through this API under these licenses. See https://www.mediawiki.org/wiki/REST_API for background and details.

Endpoint documentation

Please consult each endpoint's documentation for details on:

- Licensing information for the specific type of content and data served via the endpoint.
- Stability markers to inform you about development status and change policy, according to our API version policy.
- Endpoint specific usage limits.

Terms of service

the Wikimedia Services team - Website

Software available under the Apache 2 license

Math formula rendering

Pageviews data

Unique devices data

Legacy data

Edited pages data

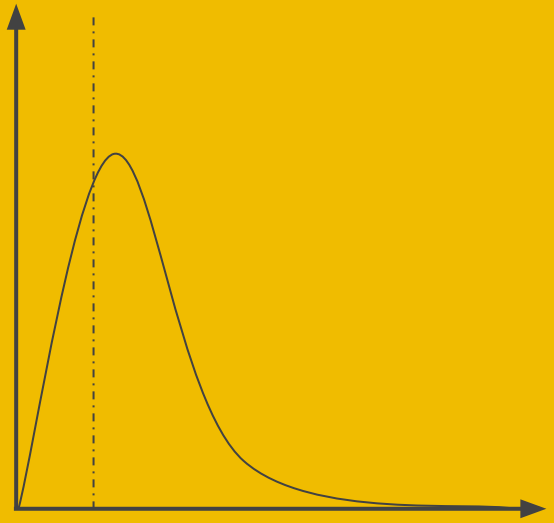
Editors data

Edits data

Existing privacy methods

Filtering

+



Thresholding

+

Bucketing

WMF's Lean Data Diet

Defined by our Privacy Policy and Data Retention Guidelines:

- No account required to read or edit
- No tracking cookies
    - Hash device IP address and UserAgent to get "Actor Signature"
- No saving data forever
    - Almost all data is aggregated/anonymized and deleted 90 days after collection

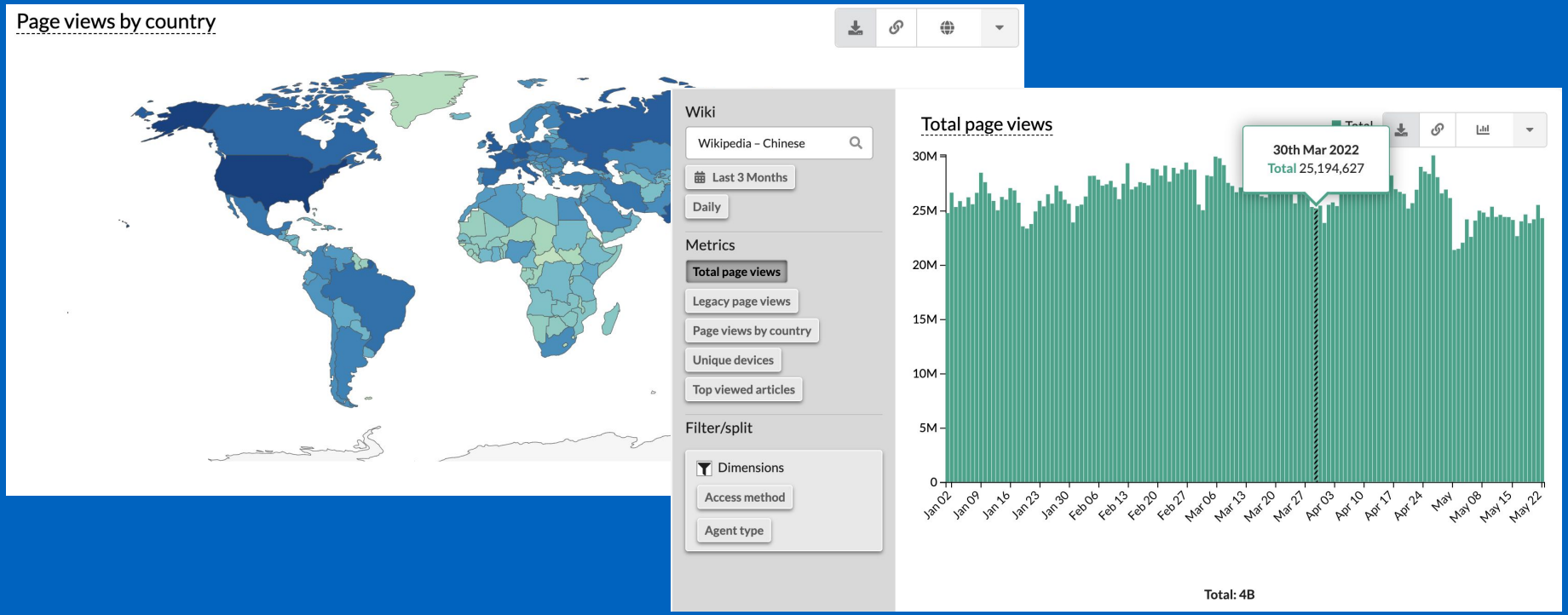
Tension between privacy and transparency

Tension between privacy and transparency ⇒ DP could be useful

02

Differential Privacy Pilot

The problem


Page views by country
Wiki
Wikipedia – Chinese
Last 3 Months
Daily
Metrics
Total page views
Legacy page views
Page views by country
Unique devices
Top viewed articles
Filter/split
Dimensions
Access method
Agent type
Total page views
30th Mar 2022
Total 25,194,627
30M
25M
20M
15M
10M
5M
0
Jan 02
Jan 09
Jan 16
Jan 23
Jan 30
Feb 06
Feb 13
Feb 20
Feb 27
Mar 06
Mar 13
Mar 20
Mar 27
Apr 03
Apr 10
Apr 17
Apr 24
May
May 08
May 15
May 22
Total: 4B

Can we use DP to release pageview counts by both project *and* country?

Country-project-page release

Basic approach:

1. For each user, define number of views per page and pages
2. Truncate dataset
3. Group-by country-project-page and sum

Country-project-page release

Country-project-page release

Install and test Tumult Labs analytics software	✅
Implement naive version of the country-project-page algorithm	✅
Define error metrics	✅
Refine parameters for naive version of algorithm through experiments	✅
Implement client-side filtering mechanism	🔄
Implement revised version of country-project-page algorithm	🔄
Productionize and automate algorithm pipeline	🔄

Country-project-page release

Why is this a ***useful*** problem to solve?

- Disaggregate trends within languages that are spoken in many countries
  - Spanish, English, Arabic, Vietnamese, Chinese, etc.
- Largest (and most unwieldy) dataset that WMF has
  - If we can successfully do it here, we can do it anywhere

Country-project-page release

Why is this a ***difficult*** problem to solve?

- High cost of failure
  - Censorship, sensitive topics, unmasking of editors, etc.
- Many country-project combos identify small user groups
- Minimizing data collection conflicts with DP

Need to do DP *carefully*

Need to do DP *differently*

Country-project-page release

Why is this a ***difficult*** problem to solve?

- High cost of failure
  - Censorship, sensitive topics, unmasking of editors, etc.
- Many country-project combos identify small user groups
- **Minimizing data collection conflicts with DP**

Need to do DP *carefully*

Need to do DP *differently*

Fundamental tension between data minimization and DP

Minimizing data collection impedes defining a strong, meaningful, and explainable notion of privacy protection for use in DP

Fundamental tension

What is a “user”?

Failure 1: One user, many signatures

IP address changes while browsing, so signature changes as well. Registered in WMF system as multiple people.

Failure 2: Many users, one signature

Many users have same IP and UA, so they all hash to the same signature. Registered in the WMF system as one person.

Fundamental tension

Failure 1: One user, many signatures

Linearly degrades privacy guarantees of DP to the extent that a user might switch IP addresses.

Meaningful issue for areas where most browsing happens on mobile (India, Indonesia, Mexico, etc.)

Failure 2: Many users, one signature

Data that could be included in count is unnecessarily dropped.

Meaningful issue for browsing within institutions where people might all have the same devices (universities, offices, etc.)

Fundamental tension

Why not just implement first-party tracking cookies?

- We **do not want to know** that data from two distinct devices, browsers, or networks comes from the same user
- This principle is **fundamentally in tension** with a system that bounds contributions from each user across all devices, browsers, and networks
- Cross-device user matching and device fingerprinting are well-researched areas — we are **deliberately choosing not to implement** that research

Fundamental tension

Data releases (like all code) *encode values*

- Stated values conflict with a system that provides precise privacy guarantees
- Unlinkability and minimizing data collection > precise privacy guarantees

We can still do better than *just* `ActorSignature`, though

Anonymous client-side filtering

Anonymous client-side filtering

Goal: A cookie attached to each web request that tells WMF whether or not that page should be included in the differentially-private aggregation for the day, up to a certain threshold k.

Anonymous client-side filtering

Failure 1: One user, many signatures

Stability > `ActorSignature`, because cookies are cleared and browser changes less than IP address changes

Failure 2: Many users, one signature

Disaggregation is possible, because distinct devices will all say to include their first *k* pages.

Anonymous client-side filtering

Implementation sketch:

```
cookie = []
salt = <global random string on server, regenerated daily>
upon pageview:
    code = md5(url, salt)[:3]
    for i in len(cookie):
        code = md5(code)
    if code not in cookie:
        for i, c in enumerate(cookie):
            cookie[i] = md5(c)
        cookie.append(code)
```

Anonymous client-side filtering

Strengths:

- Daily-rotated global salts
  - Server access ≠ decoding pageviews
  - Salt expires at midnight UTC → no connections across days
- Rehashing of cookies upon each pageview → no connection across views
- 3 character (hex) fingerprint → 4,096 combinations
  - For 10 pageviews, only ~1.1% chance of collisions within cookie
  - For English Wikipedia, any hashcode could refer to ~1,500 distinct pages

03 Open questions

Open questions (re: anonymous client-side filtering)

How to communicate these concepts with a wide audience that is highly privacy conscious?

Does anonymous client-side filtering provide a strong-enough privacy guarantee?

Difficult to test the efficiency of this methodology without compromising user privacy

Open questions (re: DP generally)

How do we continuously monitor pipeline output metrics and address any data drift that occurs?

Open questions (re: DP generally)

How do we educate stakeholders (e.g., editors) — some of whom could be ostracized, penalized, or prosecuted because of what they read/write on Wikipedia — on the purpose, scope, and protections of differential privacy?

Given this context, how do we configure our algorithms — i.e. set epsilon, delta, sensitivity, and release threshold correctly — appropriately and with informed community input?

Thank you. Questions?