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

Hal Triedman
23 June 2022
Table of contents

1. Introduction + context
2. WMF’s pilot project
   a. Problem description
   b. Fundamental tension between data minimization and DP
   c. Anonymous client-side filtering
3. Open questions
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
## Monthly Overview

### Reading

<table>
<thead>
<tr>
<th>Total page views</th>
<th>22B</th>
<th>April ↓ -4.82% month over month</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>262B</td>
<td>April ↓ -6.55% year over year</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Last 12 Months (May 2021 - Apr 2022)</td>
</tr>
</tbody>
</table>

### Page views by country

<table>
<thead>
<tr>
<th>Countries with the most views for April</th>
<th>Page views</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States of America</td>
<td>3B</td>
</tr>
<tr>
<td>Japan</td>
<td>987M</td>
</tr>
<tr>
<td>Germany</td>
<td>885M</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>796M</td>
</tr>
</tbody>
</table>

### Contributing

<table>
<thead>
<tr>
<th>Edits</th>
<th>46M</th>
<th>April ↓ -11.31% month over month</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>520M</td>
<td>April ↓ -5.34% year over year</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Last 12 Months (May 2021 - Apr 2022)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>New registered users</th>
<th>237K</th>
<th>April ↓ -11.02% month over month</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3M</td>
<td>April ↓ -18.74% year over year</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Last 12 Months (May 2021 - Apr 2022)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User edits</th>
<th>26M</th>
<th>April ↓ -9.31% month over month</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>287M</td>
<td>April ↓ -2.08% year over year</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Last 12 Months (May 2021 - Apr 2022)</td>
</tr>
</tbody>
</table>

### Content

<table>
<thead>
<tr>
<th>Total media requests</th>
<th>71B</th>
<th>April ↓ -8.59% month over month</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>882B</td>
<td>April ↓ -14.75% year over year</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Last 12 Months (May 2021 - Apr 2022)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Net bytes difference</th>
<th>20GB</th>
<th>April ↓ -68.35% month over month</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>409GB</td>
<td>April ↓ -51.27% year over year</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Last 12 Months (May 2021 - Apr 2022)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Absolute bytes diff</th>
<th>25GB</th>
<th>April ↓ -66.42% month over month</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>491GB</td>
<td>April ↓ -49.27% year over year</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Last 12 Months (May 2021 - Apr 2022)</td>
</tr>
</tbody>
</table>
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](https://www.mediawiki.org/wiki/Privacy_Policy) and [Privacy Policy](https://www.mediawiki.org/wiki/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 GNU Free Documentation License, and you irrevocably agree to release modifications or additions made through this API under these licenses. See [https://www.mediawiki.org/wiki/REST_API](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](https://www.mediawiki.org/wiki/REST_API).
- 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
Differential Privacy Pilot
The problem
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

<table>
<thead>
<tr>
<th>Task</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Install and test Tumult Labs analytics software</td>
<td>✔️</td>
</tr>
<tr>
<td>Implement naive version of the country-project-page algorithm</td>
<td>✔️</td>
</tr>
<tr>
<td>Define error metrics</td>
<td>✔️</td>
</tr>
<tr>
<td>Refine parameters for naive version of algorithm through experiments</td>
<td>✔️</td>
</tr>
<tr>
<td>Implement client-side filtering mechanism</td>
<td>🔄</td>
</tr>
<tr>
<td>Implement revised version of country-project-page algorithm</td>
<td>🔄</td>
</tr>
<tr>
<td>Productionize and automate algorithm pipeline</td>
<td>🔄</td>
</tr>
</tbody>
</table>
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 \textit{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 \textit{carefully} \}

\{ Need to do DP \textit{differently} \}
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.
What is a “user”?

$$\text{ActorSignature} = \text{MD5(} \text{IP}, \text{UserAgent} \text{)}$$

**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
Linearily 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
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
Anonymous client-side filtering

**Failure 1: One user, many signatures**

Stability $> \text{ActorSignature}$, 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.
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?