Bot Traffic at Wikimedia:

Measurement, Identification, and Response

WIKIMEDIA
FOUNDATION

Wikimedia Foundation

- The Foundation was created in 2003 as a non-profit, two years *after* Wikipedia was founded
- It hosts the infrastructure and fundraises for the Wikimedia projects (including Wikipedia) and develops tools to empower Wikimedia communities

- **Mission**: to empower and engage people around the world to collect and develop educational content under a free license or in the public domain, and to disseminate it effectively and globally.

WIKIMEDIA
FOUNDATION

Birgit Müller
Director of Product
MediaWiki & Developer Experiences
Berlin, Germany

Chris Petrillo
Principal Product Manager
Wikimedia Enterprise
Seattle, USA

Multi-generational

How can we make Wikipedia a sustainable *multi-generational* project, to serve the “in perpetuity” part of our mission?

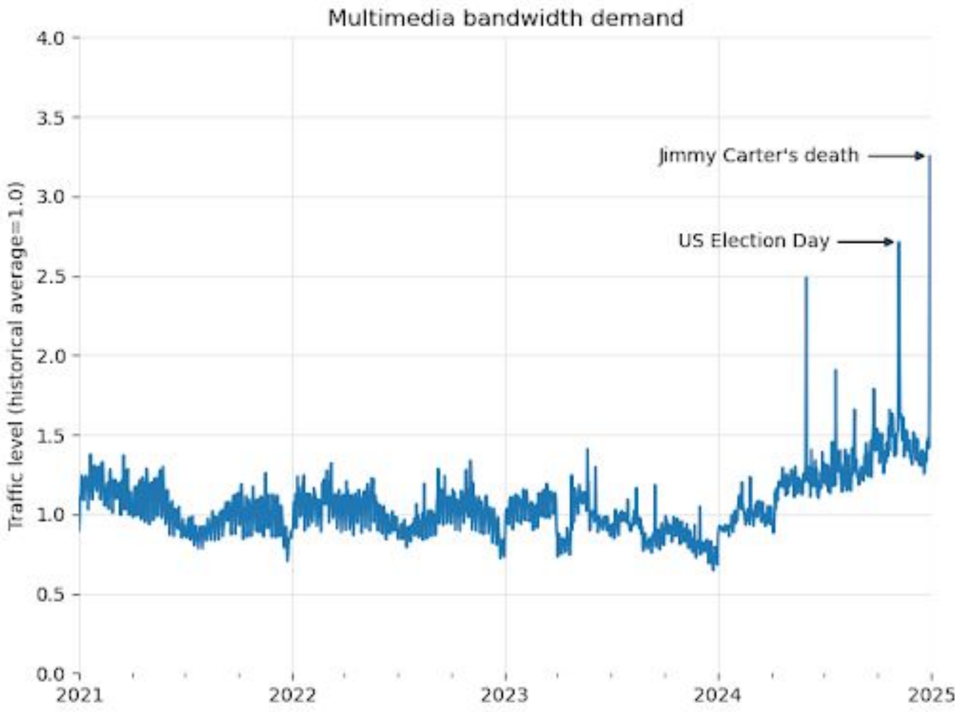
NASA/GSFC/Reto Stöckli, Nazmi El Saleous, and Marit Jentoft-Nilsen, Public domain, via Wikimedia Commons

The Problem with Bots

- Since January 2024, the Wikimedia Foundation has seen a 50% increase in bandwidth usage, largely caused by bots.

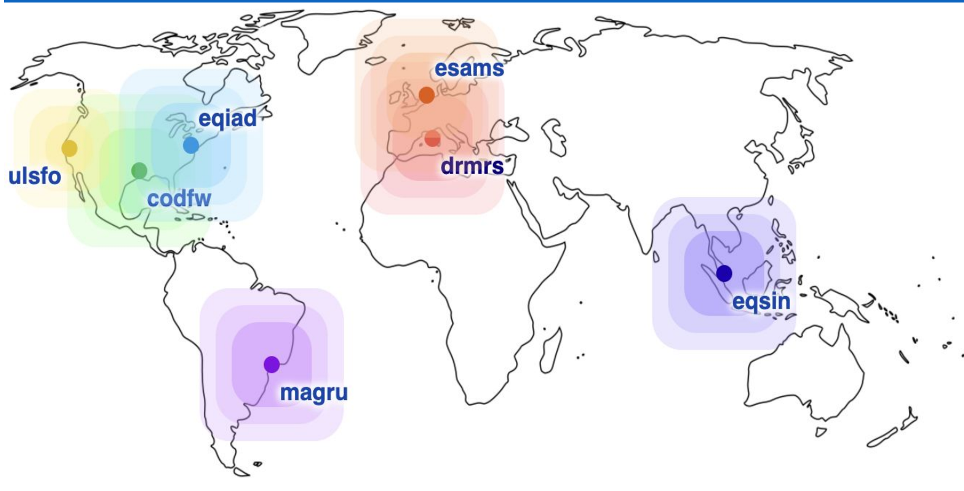
- Crawlers target not only the Wikimedia projects, but any URL within our infrastructure (e.g.: bug tracker, code review platform).

- Resources to host this demand have increased, especially given the newfound interest in high volumes of multimedia content.



Content delivery network: Optimized for humans

- 2 primary data centers, 5 caching centers
- Location-based DC-routing
- About 25B page views monthly
- Can sustain sudden, huge traffic spikes
- Requests by humans frequently terminate in the cache
- Requests by bots are more likely to get passed on to the main Data Center -> crawlers visit any URL; including less popular pages

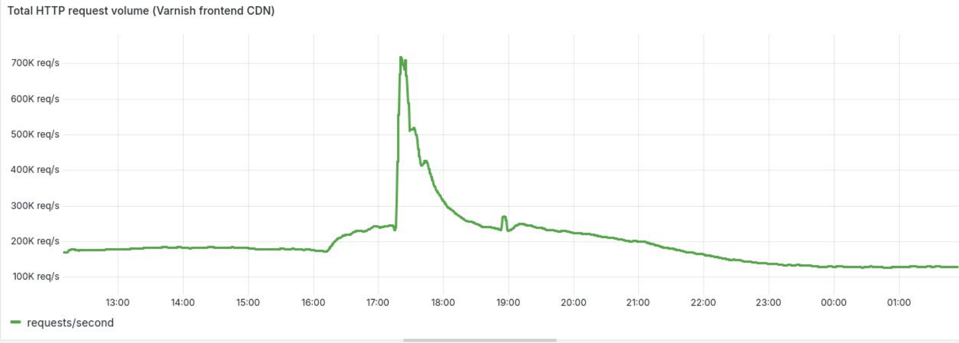


https://wikitech.wikimedia.org/wiki/Template:ClusterMap

When a big event occurs:

- Traffic spikes as everyone links to Wikipedia to read about the individual/event
- At the same time, editors frequently update the page with new information
- This results in read traffic focusing on an article that is constantly being invalidated from caches

Traffic spike when Pope Leo XIV was announced on May 8th, 2025:



Bot traffic is expensive

- 65% of our most expensive traffic comes from bots using all available access points
- 35% of our pageviews come from bots crawling all domains and projects
- Constant disruption and high workload for teams who take care of infrastructure

Whac-A-Mole_Pier_45 by Taurus Emerald, [CC-BY SA 2.0 via Wikimedia Commons]

Attribution is key

- Our mission relies on the people to find us and join as readers, contributors and donors
- Attribution is essential to ensure credibility of information and to recognise the work of our volunteers
- Adhering to the requirements of content licenses during reuse is key to enable that – no matter whether the access method is scraping, dumps, APIs, or other mechanisms

Our challenge:

Our content is free, our infrastructure is not

- Wikipedia is ***open***. Putting it behind a wall is opposed to our mission to provide the world access to knowledge
- It is hard to distinguish one robot user from another. Anyone can pretend to be “AIbot”. Mitigating mechanisms need care. Responses to limit problematic behaviour can impact all reusers not just those causing issues.
- We encourage reuse of Wikipedia but in ways that supports its sustainability.

Our challenge:

Our content is free, our infrastructure is not

- We expect reusers to support Wikimedia infrastructure based on their use.
    a. Heavy reusers can use Wikimedia Enterprise to gain access specific to their needs and to help sustain Wikimedia projects.
    b. Reusers should cite Wikimedia source links and use our branding guidelines.
    c. Reusers can partner with us to improve community growth.
- Our mission relies on people to find us and join as readers, contributors and donors

Changes in bot behaviour

- Crawling our sites for content is not new. Researchers have always worked with content, as have companies for other uses like search.

- Sheer amount of traffic generated by crawlers causes a strain on our infrastructure, combined with patterns that include a high frequency of requests, lack of attribution, and spoofing identities.

Changes in bot behaviour - Learnings

- Big companies don't always have insight into traffic due to organizational size/scope. Multiple teams may take the same data.

- Some ignore established best practices, crawling web links explicitly excluded from crawling; not honoring directives in robots.txt; and circumventing rate-limits.

- New players don't understand damage and cost associated with use, or support needed to sustain the Commons they rely on as "free data" outside of immediate short term product objectives.

WIKIMEDIA
FOUNDATION

A sustainable “Knowledge as a Service” model

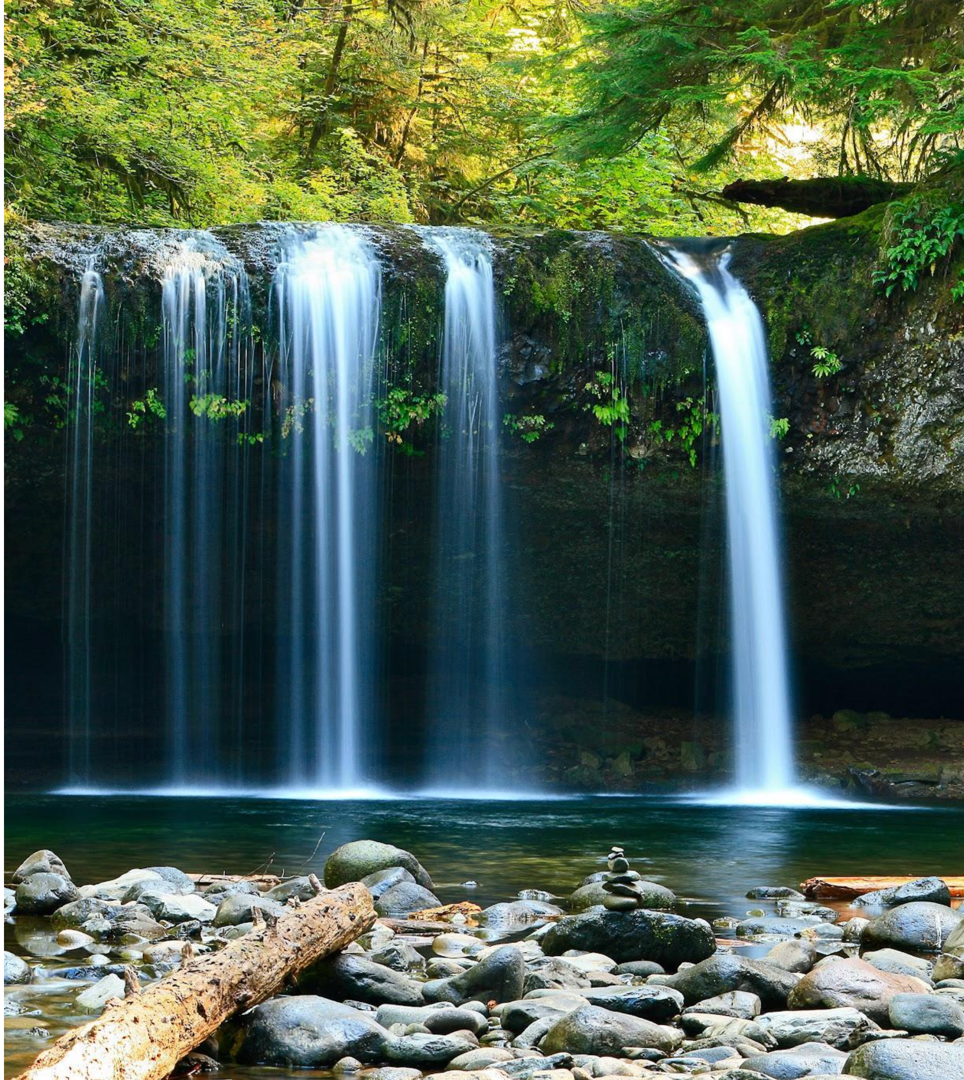
WIKIMEDIA
FOUNDATION

What that means is ...

- A systemic, but dynamic approach, enabling us to maintain preferential access for human and mission-oriented traffic and responding to changes

- Investment in our API ecosystem to meet current and future users needs and to migrate high volume users to better access methods

- Enable a base level of governance across API and scraping traffic, including tiered levels of access and the ability to enforce limits systemically

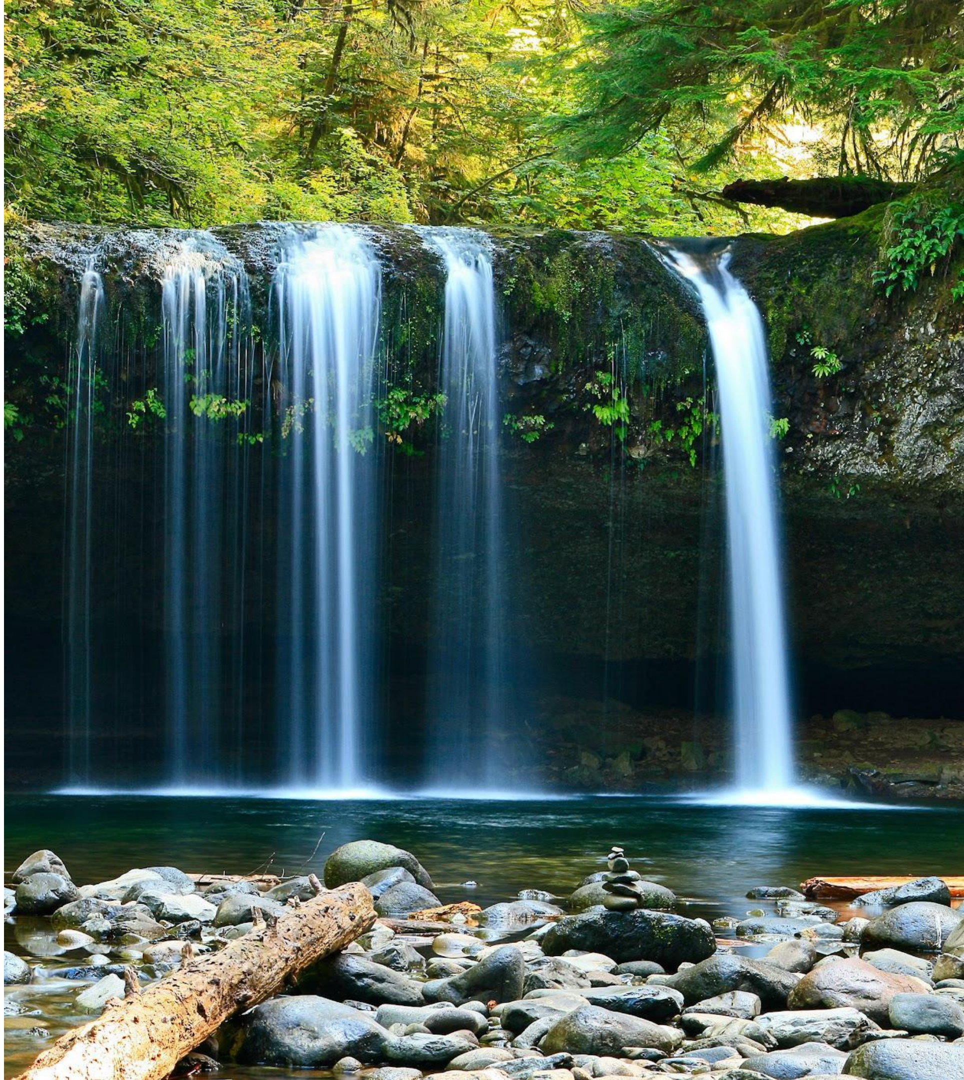


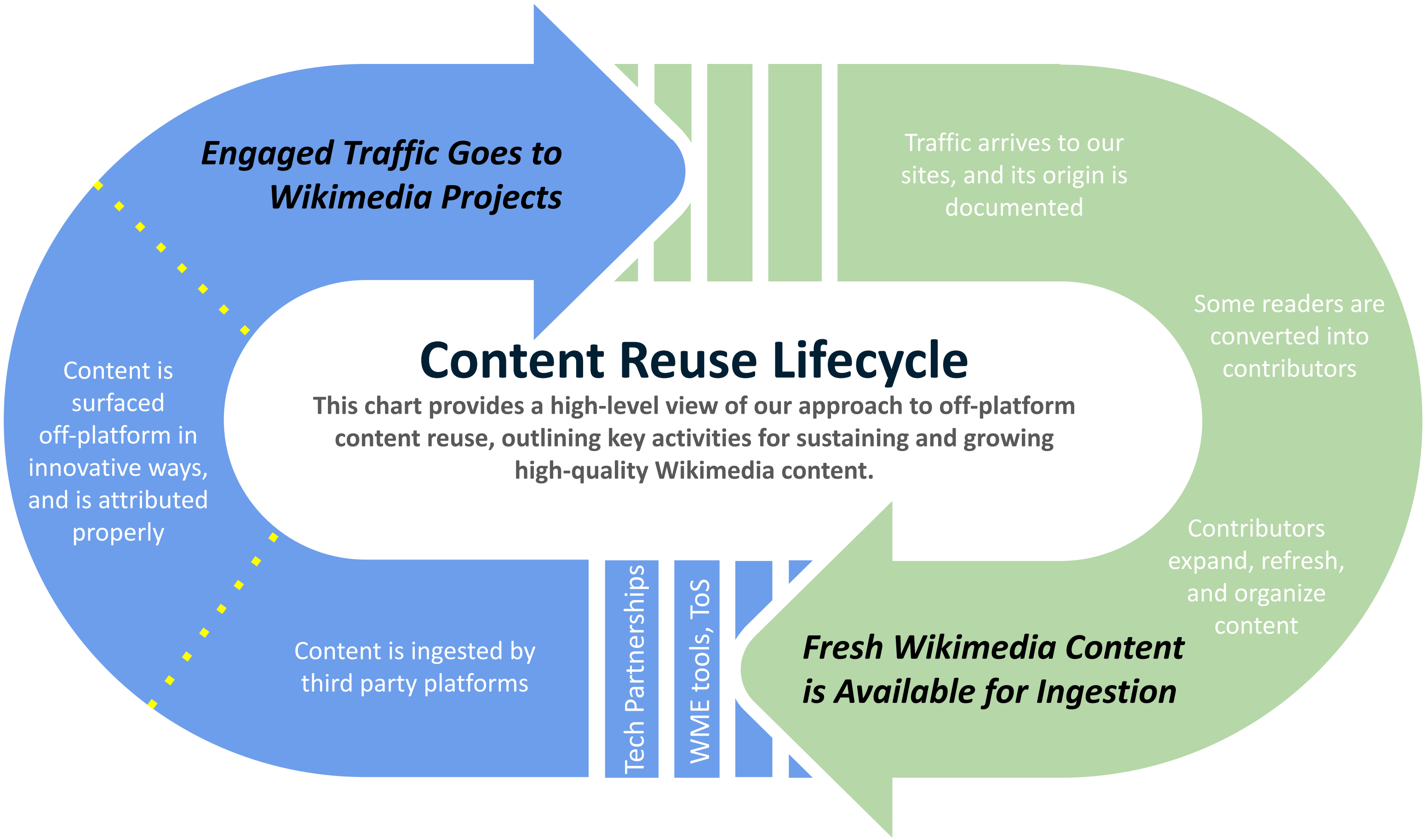
What that means is ...

- Cohesive approach to attribution for companies to help bring users back to support the resource they depend on

- Ensure adherence to our longstanding policies and guidelines (terms of use, User Agent policy, API policy, Robots policy, robots.txt, etc.)

- Enabling sharing of knowledge freely in ways that are sustainable for generations to come.




Content Reuse Lifecycle
This chart provides a high-level view of our approach to off-platform content reuse, outlining key activities for sustaining and growing high-quality Wikimedia content.
Engaged Traffic Goes to Wikimedia Projects
Traffic arrives to our sites, and its origin is documented
Some readers are converted into contributors
Contributors expand, refresh, and organize content
Fresh Wikimedia Content is Available for Ingestion
WME tools, ToS
Tech Partnerships
Content is ingested by third party platforms
Content is surfaced off-platform in innovative ways, and is attributed properly

Questions?

WIKIMEDIA
FOUNDATION