Implementing web analytics on www.ietf.org

2019-09-23

1. Introduction

Currently, no accurate data is collected about the usage of the www.ietf.org website, resulting in a lack of information upon which to base improvements to the website. Prior to the migration of the website to a CDN in May 2014, a range of usage metrics were collected and published at https://www.ietf.org/usagedata/ using web server logs and the Webalizer tool [1]. While these are still published, they are no longer accurate.

This document outlines a proposal to address this gap by implementing website analytics on the portions of the www.ietf.org website administered via the Wagtail CMS. This proposal would replace the current approach to collecting and reporting usage information.

1.1 Motivating Use Cases

Motivating use cases for these analytics include:

- *UC-1 – Understanding the visitors*
  Who are the visitors? How did visitors reach the site? How did the visitors use the site?

- *UC-2 -- How and which content is being used*
  How many visits does the site receive? What content and pages are used? When is the site used?

- *UC-3 -- What is the user experience*
  What clients and devices are being used? How quickly are the pages loading?

1.2 Design Considerations

To be consistent with the existing practices of the IETF and the specific requirements developed for the latest redesign of www.ietf.org, website analytics must be implemented to:

- limit collection and retention of data to what is needed to serve specific identified purposes;
- be documented, as appropriate, in the privacy policy;
- not require the use of web cookies; and
- not impede the use of the website via browsers that do not have JavaScript enabled.

Furthermore, the details of how analytics are implemented and how that data is used to improve the website should be available for the IETF community and any visitors for review.

2. Proposal

2.1 Scope

This proposal is scoped to include only content on www.ietf.org that is maintained in the Wagtail CMS [2]. Insights gained from this implementation may motivate consideration for wider, albeit tailored, deployments on sites such as datatracker.ietf.org and mailarchive.ietf.org. However, only www.ietf.org is in scope of this proposal.

2.2 Technology Selection

After considering several options for implementing analytics, the choices were narrowed to self-hosted analytics packages. After consulting with the IETF Tools Team, the proposal is to deploy “django-analytical” [3] application in conjunction with the Matomo On-Premise [4] analytics package.

2.3 Data Collection

While Matomo provides a broad range of functionality through the Matomo (Piwik) JavaScript Tracker [5], only a limited subset will be used. Table 1 summarizes the planned data collection, the motivation for collecting each field (per the use cases in Section 1.1) and whether this collection occurred in the current Webalizer-based system.

Table 1: Planned Data Collection

Item	Description	Previously Collected ?	UC-1	UC-2	UC-3
IP addresses	IP addresses of the request	Yes	X		
Timestamp	Approximate data and time a site resource was requested	Yes	X	X	X
Page Title	The title of the requested web page (from the HTML <title> tag)	No	X	X	
Page URL	URL of the requested resource	Yes	X	X	
Referer URL	URL of the page that linked to the requested resource	Yes	X		
Files Downloads	Identifies which non-HTML resources were downloaded from the current page. See Matomo Download documentation[7].	Yes		X	
Outside Link Clicks	Identifies which links to sites outside www.ietf.org were clicked on the current page. See Matomo Outlink documentation[8].	No	X	X	
Page Speed	Track the time it takes for web pages to be generated by the webserver and then downloaded by the requestor. See Matomo Page Speed documentation[9].	No			X
Browser Language	The preferred language of the requestor's browser (derived from the HTTP Accept-Language header)	No	X		X
User Agent	The user agent string of the browser making the request (derived from the HTTP User-Agent header). See Matomo (Piwik) Universal Device Detection Library documentation[10].	Yes			X

3. Security Considerations

3.1 Operational Security

Steps to ensure the security of the data and infrastructure will be reviewed and implemented based on the recommendation of the IETF Tools Team in coordination with the IETF Secretariat team, and will be consistent with current security practices for existing IETF data stores and infrastructure.

3.2 Access Control

Matomo collects the raw visitor data defined in Section 2.3, and then computes aggregate data (reports) summarizing this raw data. The analog in the current system is that the web server access logs are the raw visitor data and the Webalizer reports are the aggregate data.

The aggregated data will be made available to the IETF LLC staff, contractors whose role requires it, and the IESG. Providing a publicly-available summary of aggregated data will be explored and implemented if it can be done so consistently with the IETF's data security and privacy practices and policies.

Access to the raw visitor data will be restricted to only those users required to operate the system.

4. Privacy Consideration

4.1 Background

The planned configuration will only use client-side JavaScript to collect all metrics. The Matamo Image Tracker [11] feature which allows limited metric collection without JavaScript will be disabled.

A visitor can prevent all web analytics functionality by disabling JavaScript for www.ietf.org in their browser. As noted in Section 1.2, a design goal of www.ietf.org is for the website to function without JavaScript enabled.

4.1 Anonymization

The collection and reporting of these website usage metrics will entail the handling of IP addresses which in certain environments might enable user identification.

Therefore, the product will be configured to apply the “Matomo level 2” anonymization scheme [12]:

- IPv4 – mask the lower 16 bits of the address
- IPv6 – mask the lower 80 bits of the address

IP addresses will not be logged in un-anonymized form by the analytics system..

The product will also be configured to minimize the long-term re-identification of users across visits. Specifically, this will entail disabling tracking cookies [13] and not using the Matomo User ID feature [14] in the Tracking API [15] which allows for persistent user identification (even across networks).

Returning visitor statistics (i.e., the linking of multiple page requests) will be enabled based on dynamically calculated fingerprint that uses the “operating system, browser, browser plugins, [anonymized] IP address and browser language” [16]. The lifetime of this fingerprint will be 30 minutes [17].

Even with this proposed configuration, there is residual risk in the above approach that could lead to the identification of users:

- Geolocation of these IP addresses (in concert with the Browser Language) is an expected analysis. For countries with small number of IETF participants, one might be able to infer their usage.
- With holistic access to the raw visitor data (likely through SQL-level access to the underlying Matomo database as this is not a product feature), novel de-anonymization approaches could be possible. This risk will be mitigated by restricting access to the database (and raw visitor information) per Section 3.2.

4.2 Retention

The product will initially be configured with data retention periods defined in Table 2. Data beyond this period will be purged.

Table 2: Retention Periods

Data Set	Planned System (Matomo)	Current System (Webalizer)
Raw Visitor Information	5 days	5 days
Aggregate Data	12 months	12 months

4.3 Policy Review

Once the final implementation details are defined, this configuration will be subject to:

- Review for GDPR compliance by IETF LLC Counsel, and
- Review for compliance with the IETF Privacy Statement [18].

5. Implementation

In addition to installing and running the Matomo On-Premise package, a modest amount of development would be needed to integrate Matomo into the IETF's Wagtail installation. A specific implementation plan will be developed after this outline proposal is finalized and approved.

Following finalization and implementation of the proposal, and beyond adjustments that become immediately apparent to ensure expected operation, the web analytics and reports will be reviewed by the IETF Tools Team and the IESG after one-year to confirm they are delivering anticipated results.

---

[1] http://www.webalizer.org/
[2] Wagtail CMS is used to maintain files in the following paths under www.ietf.org : /about, /blog, /chairs, /contact, /how, /iesg, /live, /logo, /privacy-statement, /standards, /topics, /trademark-list, /bibliography, /documents, /forms, /links, /search
[3] https://github.com/jazzband/django-analytical
[4] https://matomo.org/what-is-on-premise/
[5] https://developer.matomo.org/api-reference/tracking-javascript
[6] Whether this data item was previously reporting through the Webalizer tool.
[7] https://matomo.org/faq/new-to-piwik/faq_47/
[8] https://matomo.org/faq/new-to-piwik/faq_71/
[9] https://matomo.org/docs/page-speed/
[10] https://github.com/piwik/device-detector
[11] https://matomo.org/docs/tracking-api/#image-tracker-code
[12] https://github.com/matomo-org/matomo/commit/c072a0e5911b544890a39e86c9efe024bb2475
[13] https://matomo.org/faq/general/faq_146/
[14] https://matomo.org/docs/user-id/
[15] https://developer.matomo.org/guides/tracking-introduction
[16] https://matomo.org/faq/general/faq_21418/
[17] https://matomo.org/faq/how-to/faq_190/
[18] https://www.ietf.org/privacy-statement/
