Workgroup: Network Working Group
Internet-Draft: draft-ring-analyticstxt-00
Published: 22 April 2021
Intended Status: Informational
Expires: 24 October 2021
Authors: F. Ring H. Niefeld
Offen Offen

A File Format for the Discoverable Use of Analytics

Abstract

Internet privacy has become an important feature for users of websites and services. This document proposes a way for websites and services to declare and disclose their usage of analytics and tracking software. analytics.txt aims to be an elaborate file format that describes the privacy related characteristics of analytics and tracking software in a non-biased way. An analytics.txt file is understandable for a non-technical audience, while also useful for the automated consumption by tools and software.

Discussion Venues

This note is to be removed before publishing as an RFC.

Source for this draft and an issue tracker can be found at https://github.com/offen/analyticstxt.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 24 October 2021.

Copyright Notice

Table of Contents

1. Introduction

1.1. Motivation

User tracking and the utilization of analytics software on websites has become a widely employed routine, visibly and invisibly affecting the way the user facing internet works and behaves. Yet, there is no well-defined way of accessing information about what software is being used and what kind of data it is collecting in a standardized way. Legislation can only ever cover a subset of the range of existing technological implementations, creating incentives for software to find workarounds, thus allowing them to hide their presence from users. Automated audits are limited to aspects that are possible to detect in clients, but cannot disclose other important implementation details.

1.2. Scope of this proposal

This document defines a way to specify the privacy related characteristics of analytics and tracking software. We aim for this information to be consumable both by humans as well as software. For example, search engines or browser extensions could make use of the provided data and display information to users, but it should also be simple enough to serve as information for inquiring users as is.

The file "analytics.txt" is not intended to replace the requirement for complying with existing regulations, but supposed to give insights beyond the scope of these regulations.

1.3. Definition of the term "analytics" in the scope of this document

Analytics as referred to in this document involves the collection of usage statistics in order to generate reports that can help the providers of websites and services to better understand and optimize their services towards real world user behavior. This can also include measuring different content against different groups of users. Analytics or user tracking as referred to in this document does not refer to the identification of users in order to deliver customized advertising or content across websites of any kind.

2. Conventions and Definitions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

The term "implementors" refers to the providers of services and websites that wish to use an analytics.txt file.

3. Specification

This document defines a text file format that can be used by implementors to signal information about their usage of analytics software to both users and software.

By convention, this file is called "analytics.txt". Its location and scope are described in Section 4.

This text file contains multiple fields with different values. A field contains a "name" which is the first part of a field all the way up to the colon (for example: "Author:") and follows the syntax defined for "field-name" in section 3.6.8 of [RFC5322]. Field names are case-insensitive (as per section 2.3 of [RFC5234]). The "value" comes after the field name and follows the syntax defined for "unstructured" in section 3.2.5 of [RFC5322]. The file MAY also contain blank lines and comments.

A field MUST always consist of a name and a value (for example: "Author: Jane Doe jane.doe@example.com"). Each field MUST appear on its own line. Unless specified otherwise by the field definition, multiple values MUST be chained together for a single field (for example: "Implements: gdpr, ccpa") using the "," character (%x2c). A field MAY NOT appear multiple times.

Implementors SHOULD aim for authoring an analytics.txt file that is easy to understand by non-technical audiences.

3.1. Comments

Any line beginning with the "#" (%x23) symbol MUST be interpreted as a comment. The content of the comment may contain any ASCII or Unicode characters in the %x21-7E and %x80-FFFFF ranges plus the tab (%x09) and space (%x20) characters.

Example:

```
# This is a comment
```

Implementors SHOULD make deliberate use of comments to make an analytics.txt file more accessible for non-technical audiences.

3.2. Line Separators

Every line MUST end either with a carriage return and line feed characters (CRLF / %x0D %x0A) or just a line feed character (LF / %x0A).

3.3. Extensibility

Like many other formats and protocols, this format may need to be extended over time to fit the ever-changing landscape of the Internet. Special attention is required for defining the allowed values in enumerations to ensure they are a. extendable and b. do not become obsolete too quickly.

3.4. Field Definitions

Field names are case-insensitive, yet implementors SHOULD use the capitalized style used in this document for consistency.

Field values are case-insensitive. Unless otherwise specified, implementors MUST refer to the allowed values for a field given by the specification.

3.4.1. Author

This REQUIRED field holds an OPTIONAL author name and a REQUIRED email address providing information about a person or entity responsible for maintaining the contents of the file. The field MUST contain a valid email address which shall be used for inquiries about the correctness and additions to the data provided in the file.

3.4.1.1. Example

```
Author: Jane Doe <jane.doe@example.com>
```

3.4.2. Collects

This REQUIRED multi-value field indicates which potentially privacy relevant user specific data is being collected or used in session identification or other procedures. These values MUST also be specified if a property is not persisted as-is, but stored or processed in a hashed and/or combined form.

3.4.2.1. Allowed values

3.4.2.1.1. none

No analytics data is collected at all. This value MUST NOT be used in conjunction with other values.

3.4.2.1.2. url

The URL of a visit, including its path, is collected and used. This MUST also be specified in case URLs are stripped of certain parameters or pseudonymized before being stored.

3.4.2.1.3. ip-address

The request IP address is being used.

3.4.2.1.4. geo-location

Geographic location of users is determined and used. This could for example be derived from the request IP, or from using browser APIs.

3.4.2.1.5. user-agent

Information about the utilized User Agent is being collected.

3.4.2.1.6. fingerprint

Browser Fingerprinting is used. Such mechanisms usually try to compute a unique identifier from properties of the host Operating System, allowing them to re-identify users without having to persist an identifier.

3.4.2.1.7. device-type

The user's device type (e.g. mobile / tablet / desktop) is being determined and collected.

3.4.2.1.8. referrer

The Referrer of a visit is collected and used. This MUST also be specified if the referrer value is stripped of potential path fragments.

3.4.2.1.9. visit-duration

The duration of a visit, either on page- or on session-level is measured and used.

3.4.2.1.10. custom-events

Custom events like conversion goals are defined and used. This MAY be left out in case the analytics software in use offers such functionality, but implementors chose not to use the feature.

3.4.2.1.11. session-recording

Detailed behavior like mouse movement and scrolling is recorded and can possibly be played back when analyzing the analytics data.

3.4.2.2. Example

Collects: url, device-type, referrer

3.4.3. Stores

This field is REQUIRED unless the only value of the Collects field as per Section 3.4.2 is none. The multi-value field indicates whether data is persisted on the client during the collection of analytics data and declares the browser features used for doing so. In case no data is being persisted at all, the value none MUST be used as the single entry for this field.

3.4.3.1. Allowed values

3.4.3.1.1. none

No data is persisted on the client during the collection of usage data. This value MUST NOT be used in conjunction with other values.

3.4.3.1.2. first-party-cookies

First party cookies are in use. There is no differentiation between session or persistent cookies, just like HTTP and JavaScript cookies are considered equal.

3.4.3.1.3. third-party-cookies

Third party cookies are in use. There is no differentiation between session or persistent cookies, just like HTTP and JavaScript cookies are considered equal.

3.4.3.1.4. local-storage

Data is persisted on the client using non-cookie JavaScript APIs like localStorage, sessionStorage, WebSQL or IndexedDB

3.4.3.1.5. cache

The analytics software leverages browser caches to store identifiers. For example, ETag headers can be used to identify users based on their browser caches' contents. This value is not required in case the analytics software sends static resources with cache headers, but does not make use of the request headers on subsequent requests.

3.4.3.2. Example

Stores: first-party-cookies, local-storage

3.4.4. Uses

This field is REQUIRED unless the only value of the Collects field Section 3.4.2 is none. The multi-value field indicates the technical implementation details for how analytics data is being collected.

3.4.4.1. Allowed values

3.4.4.1.1. javascript

A client-side script is used to collect data.

3.4.4.1.2. pixel

A static resource - typically a pixel - transferred via HTTP is being used to collect data through the request parameters.

3.4.4.1.3. server-side

Collection of usage data is happening on the server side at application layer.

3.4.4.1.4. logs

Usage data is being calculated from server log files.

3.4.4.1.5. other

Other techniques that are not described in this section are in use.

3.4.4.2. Example

Uses: script

3.4.5. Allows

This field is REQUIRED unless the only value of the Collects field Section 3.4.2 is none. The multi-value field discloses information

about whether user consent is being acquired before collecting analytics data, and if it is possible for users to opt out of the collection of usage data.

3.4.5.1. Allowed values

3.4.5.1.1. none

The software does not define a way for users to opt in or opt out of the collection of usage data. This value also applies to scenarios where only a subset of data is collected by default and could be extended by opting in. This value MUST NOT be used in conjunction with other values.

3.4.5.1.2. opt-in

No usage data is collected before users have given their consent.

3.4.5.1.3. opt-out

Users can opt out of collection of usage data using a dedicated feature tailored towards the user audience. This value is only applicable in case no data at all is collected after having opted out.

3.4.5.2. Example

```
Allows: opt-out
```

3.4.6. Retains

This field is REQUIRED unless the only value of the Collects field Section 3.4.2 is none. The single-value field indicates the duration for which the analytics data is being stored before being deleted. The value is either a duration as defined in [RFC3339] or the token "perpetual" in case data is retained without expiring it at some point. Implementors SHOULD add a comment providing a human readable value to this field.

3.4.6.1. Example

```
# Data is retained for twelve months
Retains: P12M
```

3.4.7. Honors

This OPTIONAL, RECOMMENDED multi-value field indicates which browser level privacy controls are being honored when collecting data.

3.4.7.1. Allowed values

3.4.7.1.1. none

Data is collected even if any of the browser settings listed below are in use. This value MUST NOT be used in conjunction with other values.

3.4.7.1.2. do-not-track

User-Agents that have DoNotTrack [DNT] enabled will be excluded from the collection of analytics data.

3.4.7.1.3. global-privacy-control

User agents that have Global Privacy Control [GPC] enabled will be excluded from the collection of analytics data.

3.4.7.2. Example

Honors: do-not-track, global-privacy-control

3.4.8. Tracks

This OPTIONAL, RECOMMENDED multi-value field indicates the coverage in session and user lifecycle tracking.

3.4.8.1. Allowed values

3.4.8.1.1. none

Each event that is collected is anonymous. There is no way to connect and group multiple pageviews by user or similar. This value MUST NOT be used in conjunction with other values.

3.4.8.1.2. sessions

Metrics that source from a single browser session can be grouped and distinguished as such.

3.4.8.1.3. users

Users can be identified across multiple browser sessions.

3.4.8.2. Example

Tracks: sessions, users

3.4.9. Varies

This OPTIONAL, RECOMMENDED single-value field indicates the usage of content experiments like A/B testing. It MUST contain a single value only.

3.4.9.1. Allowed values

3.4.9.1.1. none

All users are served the same content without any changes. This value MUST NOT be used in conjunction with other values.

3.4.9.1.2. random

Content experiments are performed by grouping users randomly into buckets and serving them different content.

3.4.9.1.3. behavioral

Content experiments are performed by grouping users into buckets based on their behavior and serving them different content.

3.4.9.2. Example

Varies: random

3.4.10. Shares

This OPTIONAL, RECOMMENDED multi-value field indicates whether data is shared with select users, the general public or third parties.

3.4.10.1. Allowed values

3.4.10.1.1. none

The data collected is not shared with any party unless directly affiliated with the implementor, e.g. employees.

3.4.10.1.2. per-user

Users can access the usage data that is associated with them in a non-aggregated way, isolating all data that is specific to their current means of re-identification.

3.4.10.1.3. general-public

Usage statistics for the site or service are available to the general public.

3.4.10.1.4. third-party

Data is being shared non-publicly with third parties. This MUST also be specified when datasets are aggregated or pseudonymized beforehand.

3.4.10.2. Example

```
Shares: general-public
```

3.4.11. Implements

This OPTIONAL field indicates conformance with existing regulations and legislation. Values for this field SHOULD use all lowercase tokens with whitespace being replaced by the dash character (%x2d). This field SHOULD only be added if it makes the setup described by the file easier to understand for human users.

Example values are:

- *gdpr
- *ccpa

3.4.11.1. Example

```
Implements: gdpr, ccpa
```

3.4.12. Deploys

This OPTIONAL field indicates which software is being used for collecting analytics. Values for this field SHOULD use all lowercase tokens with whitespace being replaced by the dash character (%x2d). This field SHOULD only be added if it makes the setup described by the file easier to understand for human users.

Example values are:

- *google-analytics
- *plausible
- *hotjar
- *matomo

3.4.12.1. Example

```
Deploys: google-analytics, hotjar
```

3.5. Examples of analytics.txt files

3.5.1. A site using analytics

```
# analytics.txt file for www.example.com
Author: Jane Doe <doe@example.com>

Collects: url, referrer, device-type
Stores: first-party-cookies, local-storage
# Usage data is encrypted end-to-end
Uses: javascript
# Users can also delete their usage data only without opting out
Allows: opt-in, opt-out
# Data is retained for 6 months
Retains: P6M

# Optional fields
Honors: none
Tracks: sessions, users
Varies: none
Shares: per-user
Implements: gdpr
```

3.5.2. Specifying required fields only

```
Author: John Doe <doe@example.com>
Collects: url, ip-address, geo-location, user-agent, referrer, device-ty
Stores: none
Uses: javascript
Allows: none
Retains: perpetual
```

3.5.3. A site not using any analytics

```
# analytics.txt file for www.example.com
Author: Jane Doe <doe@example.com>
Collects: none
```

4. Location of the analytics.txt file

By default, an analytics.txt file SHOULD be placed in the ".well-known" path as per [RFC8615] of a domain name or IP address.

4.1. Alternatives

In case implementors are unable to meet this requirement, other options are available.

4.1.1. link Tag

Implementors MAY signal the location of an analytics.txt file in the context of a HTML document using a link element of rel "analytics"

Example:

```
<link rel="analytics" href="https://example.com/resources/analytics.txt"
```

4.1.2. HTTP Header

Implementors MAY send an HTTP header of X-Analytics-Txt with a response, sending the URI of the applicable file.

Example:

```
X-Analytics-Txt: https://example.com/resources/analytics.txt
```

4.2. Precedence

In case multiple of these signals are being used, the precedence taken is:

1. X-Analytics-Txt Header
2. link element
3. ".well-known" location

4.3. Scope of a file

An analytics.txt file located in the ".well-known" location MUST only apply to the domain or IP address of the URI used to retrieve it, and SHALL NOT apply to any of its subdomains or parent domains. If the location is signaled using the HTTP Header or in the document markup itself, its scope SHALL be limited to the requested resource only.

If distributed in non-standard locations, an analytics.txt file MAY also apply to products and services provided by the organization publishing the file (e.g. desktop or mobile applications) and which cannot be mapped to a domain name or IP address. In such cases, implementors MUST add sufficient commentary describing the applicable scope.

5. Security Considerations

5.1. Incorrect or stale information

If information given in an "analytics.txt" file is incorrect or not kept up to date, this can result in usage of services under wrong assumptions, thus exposing users to possibly unwanted data collection and handling. Not having an "analytics.txt" file may be preferable to having incorrect or stale information in this file. This guideline also applies to field level: in case of ambiguities or uncertainties, it's recommended to omit a field or a value rather than providing incorrect information. Implementors MUST use the "Author" field (see Section 3.4.1) to allow inquiries about the correctness of the given information.

5.2. Spam

Implementors should be aware that disclosing mandatory author information as per Section 3.4.1 in such a file exposes them to possible Spam schemes or spurious requests.

5.3. Multi-user environments

In multi-user / multi-tenant environments, it may possible for a single user to take over the location of the "/.well-known/security.txt" file which would also apply to others. Organizations should ensure the ".well-known" location is properly protected. Implementors can instead use other locations as per Section 4 in such scenarios.

6. IANA Considerations

6.1. Well-Known URIs registry

The "Well-Known URIs" registry should be updated with the following additional values (using the template from [RFC8615]):

URI suffix: analytics.txt

Specification document(s): this document

Status: permanent

7. References

7.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://doi.org/10.17487/RFC2119>.

[RFC3339] Klyne, G. and C. Newman, "Date and Time on the Internet: Timestamps", RFC 3339, DOI 10.17487/RFC3339, July 2002, <https://doi.org/10.17487/RFC3339>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://doi.org/10.17487/RFC8174>.

[RFC8615] Nottingham, M., "Well-Known Uniform Resource Identifiers (URIs)", RFC 8615, DOI 10.17487/RFC8615, May 2019, <https://doi.org/10.17487/RFC8615>.

7.2. Informative References

[DNT] Fielding, R.T. and D. Singer, "Tracking Preference Expression (DNT)", n.d., <https://www.w3.org/TR/tracking-dnt/>.

[GPC] Berjon, R., Zimmeck, S., Soltani, A., Harbage, D., and P. Snyder, "Global Privacy Control (GPC)", n.d., <https://globalprivacycontrol.github.io/gpc-spec/>.

[RFC5234] Crocker, D., Ed. and P. Overell, "Augmented BNF for Syntax Specifications: ABNF", STD 68, RFC 5234, DOI 10.17487/RFC5234, January 2008, <https://doi.org/10.17487/RFC5234>.

[RFC5322] Resnick, P., Ed., "Internet Message Format", RFC 5322, DOI 10.17487/RFC5322, October 2008, <https://doi.org/10.17487/RFC5322>.

Authors' Addresses

Frederik Ring
Offen

Email: frederik.ring@gmail.com

Hendrik Niefeld
Offen

Email: hello@niefeld.com