HARVARD UNIVERSITY

Enterprise Architecture and the Harvard Library

Advisory

Web Site Analytics

Authors: Greg Charest, Raoul Sevier Suzanne Wones	**Audience Level:** • Strategy Planning and EA Leader • Solution Architect and Program Manager
Version: 0.3 **Last Revised:** 03-Oct-2019 **Status:** draft **Document Type:** Single Topic Guidance	**Distribution Scope:** Harvard-wide
Workgroup Members:	**Reviewers:** Jefferson Burson Tim Murray

1. Problem Statement:

Website traffic analysis tools provide data and insights that can be used to create a better user experience for website visitors and to understand if a website is accomplishing its intended purpose. However, these tools collect a significant level of personal information and the unexamined and haphazard use of analytics can lead to the unnecessary collection of information about our users that could constitute a breach of trust. This advisory provides guidance on striking a balance between using the tools effectively and ensuring that user privacy is respected.

2. Discussion

Because Google Analytics has an 85% market share at this time it is the focus case for this discussion. The discussion and recommendations are generally applicable to most web analytics tools.

Understanding the community of users that interact with a web site allows site owners to tailor interactions with users and enhance sites based on usage patterns. Google Analytics enables the collection of data in three main areas;

- *technical data* about the user including information about the user's browser, ip address, the computer's operating system, page access dates and times and other data
- *behavioral data* including common landing pages, exit pages, frequently visited pages, length of time spent per visit, the page the user is coming from, etc.
- *locational data* derived from the IP address, generally with a city level resolution

Harvard website owners and administrators should consider the following three general topics when deploying and using web analytics:

- *Security* - Users should expect that their interactions with web sites and applications will be private and not subject to eavesdropping. Today many sites are moving to secure communication for all activities, not just high value activities such as commerce transactions.
- *Transparency* - Users should be aware of the privacy implications of using web tools. Privacy statements that make clear the collection and uses of data that may be collected directly or automatically should be easy to access and delivered in language that users can understand.
- *Records management* - Information collected about usage and users should be managed carefully and retained no longer than needed or required by rule. When data is to be purged, all related data should be deleted, including data stored in backups.

3. Recommendations:

3.1. General

3.1.1. Security

All websites should be protected using HTTPS even if they do not handle sensitive information. Aside from providing critical security and data integrity for both your websites and your users' personal information, HTTPS is a requirement for many new browser features. Security certificate can be requested via unix-help@harvard.edu.

3.1.2.Privacy Statements

A Privacy Policy statement should be prominently referenced on the entry page of a web site or web application. For sites that are designed to be entered through multiple entry points, a link to the Privacy Policy should be visible at each point. Harvard University has a general privacy statement that is an excellent model for other Harvard organizations.

3.1.3.Use of Cookies

Cookies are small files that are stored on your computer (unless you block them). Cookies are used to understand and save user preferences for future visits and compile aggregate data about site traffic and site. You should design web sites that will continue to function, at least minimally, should a user disable cookies through browser options or has opted out of the collection and use of this information through tools like the Network Advertising Initiative opt-out page.

3.1.4.Records Management

Consider what data you are collecting and for how long you need to keep it. Get into the practice of analyzing your data on a regular schedule and then figure out what you need to keep, what you can delete, and whether there are opportunities to collapse data in order to further protect the privacy of individuals while being able to measure trends over time. The purpose of collecting web traffic analytics data is to improve the effectiveness of a website. If the data is not used in making improvements to the site, stop collecting it.

Although web traffic data is not specifically addressed in the Harvard General Records Schedule, Records Management Services, a department of Harvard University Archives, provides guidance to University staff, faculty, and administrators on how to understand their responsibilities for stewarding and managing their records.

3.2. Google Analytics Specific

3.2.1. Security

Manage access the Google Analytics administrative account carefully. Google provides an account to access a console that manages configuration settings and data analysis tools. The account credentials should be managed using a standard lifecycle approach and distributed only to trusted individuals.

3.2.2.Privacy Statements

It is important to note that Google specifies a requirement for a privacy statement in its Terms of Service. When using Google Analytics, the privacy statement should include information on how to opt-out using an available browser plugin such as the Google Analytics Opt-out Addon.

3.2.3.Records Management

The Google Analytics terms of service explicitly prohibit the sending of personally identifiable information (PII) to Google, which includes (but isn't limited to) name, email address, and billing information. In addition, you should exercise caution when combining web site analytics data with other information that may result in creating a Harvard Level 3 data set with PII. This should extend to the generation of reports which could expose personal information unnecessarily.

The basic Google Analytics tracking setup does not collect very much personal data, but it does include IP address information. Google provides code to further anonymize IP addresses by overwriting a portion of the address with zeros. This code to anonymize IP addresses should be embedded in the Google Analytics html coding. Details are available here and here.

4. Appendix

Google Analytics - Tracking Code Overview
https://developers.google.com/analytics/resources/concepts/gaConceptsTrackingOverview

Harvard University Privacy Statement
https://www.harvard.edu/privacy-statement

Harvard University Additional EEA Privacy Disclosures
https://gdpr.harvard.edu/eeaprivacydisclosures

Harvard Records Management Services
https://library.harvard.edu/how-to/manage-harvard-university-records

Network Advertising Initiative opt-out
http://www.networkadvertising.org/managing/opt_out.asp

Opt-out Browser Add-on
https://tools.google.com/dlpage/gaoptout/

Google Analytics IP Anonymization
https://support.google.com/analytics/answer/2763052