ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 33 (diciembre 2015)

Bibliotecas y analítica web: una cuestión de privacidad

Ramiro Federico Uviña

CeDInCI, Universidad Nacional San Martín / ramirouvia@gmail.com

Resumen

Describe las formas de monitoreo web existentes, poniendo el foco en como estas herramientas gestionan la privacidad de los datos, con el objetivo de brindar recomendaciones en cuanto a la adopción de software de analítica web en las unidades de información. Para esto se realizó un estudio comparativo de la forma en que gestionan la privacidad dos herramientas de analítica web: Google Analytics y Piwik, concluyendo que para las bibliotecas es recomendable seleccionar la última para soslayar las borrosas políticas de privacidad de Google.

Palabras clave

Analítica web
Bibliotecas
Privacidad
Google Analytics
Piwik

Abstract

Libraries and Web Analytics: a Privacy Matte. This paper describes the current web monitoring techniques, emphasizing on how these tools handle the data privacy, with the purpose of establishing guidelines for the adoption of web analytic tools in libraries. Hereby, it presents a comparative analysis of privacy in web analytic software: Google Analytics and Piwik, concluding that libraries aware of privacy matters should use Piwik before Google Analytics, for the blurry Google privacy policies.

Keywords

Web analytics
Libraries
Privacy
Google Analytics
Piwik

Introducción

En el último decenio las publicaciones especializadas en Bibliotecología han comenzado a brindar espacio en sus números a la analítica web como herramienta útil a los sitios web de las bibliotecas. Como muestra más acabada de este creciente interés podemos mencionar dos ejemplos muy cercanos en el tiempo: los números especiales sobre este tema de *Library Technology Reports* de Julio de 2011 y de Junio de 2013. Siendo que esta publicación tiene como función primordial proveer a sus lectores análisis comparativos y características principales de tecnologías de la información, y entendiendo que por esto mismo sus intervenciones pueden llegar a tener un impacto

fundamental a la hora de tomar decisiones en los directivos de una biblioteca, ya que quienes se suscriben lo hacen para obtener justamente un análisis comparativo y conocer las tendencias en el uso de tecnologías de la información para actualizarse sobre un mercado rico y cambiante, sus recomendaciones tienen relevancia en la comunidad bibliotecaria mundial. No obstante, como podemos observar en estos números especiales, se trata la posibilidad de utilizar Google Analytics como software de analítica web, apelando a su facilidad de instalación y configuración (casi inexistentes, en caso de requerir reportes básicos). Debido a las recientes controversias sobre las políticas de privacidad de Google[1, 2] y entendiendo que las bibliotecas contienen datos sensibles[3], este trabajo pretende comparar desde un punto de vista crítico la polémica "recomendación" que *Library Technology Reports* hace a la comunidad de bibliotecarios de utilizar Google Analytics como herramienta de analítica web, más allá de los problemas de privacidad que esta empresa plantea, e introducir una alternativa *open source*, de gestión local, que permite mucho mayor control de los datos privados de los usuarios.

1. http://www.lavanguardia.com/tecnologia/20140901/54414563508/alerta-acusa-a-los-tecnicos-de-la-comision-europea-de-vivir-en-el-pasado.html

2. http://www.telam.com.ar/notas/201409/76537-para-el-presidente-de-telefonica-en-la-web-la-seguridad-es-inexistente.html

1. Monitoreo web

3. http://lj.libraryjournal.com/2014/08/people/library-freedom-fighter-zoia-horn-remembered/

Si bien las técnicas de monitoreo web son reflejadas en la mayor parte de la bibliografía relevada como "la promesa que realmente va a revolucionar la forma en que se hacen los negocios en la web" (Kaushik, 2010), y por tanto se relacionan mayormente con las empresas de negocios web, también es cierto que en los últimos tiempos algunas bibliotecas entraron en una lógica de costo – beneficio o de "return of investment" (Marek, 2011). A pesar de lo antedicho, Marek no adscribe a la postura de que todas las bibliotecas deban manejarse dentro de esa lógica, sí justifica o considera válido entender la biblioteca como un centro de recursos orientado a los usuarios, lo que permite entender la inclusión de herramientas propias de disciplinas como el *marketing* para medir el impacto de los servicios que estas instituciones proveen, así como la adecuación de los mismos a las necesidades de los usuarios. Queda claro entonces que la biblioteca tiene que ir incorporando nuevas formas de interactuar con sus usuarios para posibilitar la satisfacción de necesidades de información que presentan los usuarios de sus servicios web.

Las técnicas de monitoreo web intentan contestar cuatro preguntas fundamentales sobre los visitantes de los sitios web a los cuales acceden: ¿Qué hicieron? ¿Cómo lo hicieron? ¿Por qué lo hicieron? ¿Podrían hacerlo? La primera y la segunda de estas preguntas son medibles, en mayor o menor medida, a través de las herramientas de analítica web y sobre ellas nos enfocaremos.

Las formas de implementar un programa de monitoreo web pueden clasificarse como:

- Herramientas que leen el log del servidor
- Incorporación de código en las páginas (Page Tagging)
- Otras

1.1. Análisis de los logs del servidor

Los logs de los servidores existen desde que se implementó el primer servidor (Croll y Power, 2009). Si bien al principio se limitaban a ser utilizados como herramienta de análisis de funcionamiento para los administradores de los mismos (el departamento de informática), y solamente contenía información acerca de la solicitud que se le hacía al servidor (la fecha y hora en la que esta ocurría, la solicitud, el estatus http que devolvía, y la longitud en bytes del documento que devolvía) y desde dónde se conectaba el visitante. Más adelante, a mediados de la década comprendida entre

```
10.100.3.200 - - [28/Mar/2009:11:50:55 -0400] "GET / HTTP/1.1" 200 53785 "http://twitter.com/seanpower"

"Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US; rv:1.9.0.7)

Gecko/2009021910 Firefox/3.0.7" 125763
```

Figura 1 – Ejemplo de línea de log del servidor (Croll y Power, 2009: 114)

1990 y 2000 se incluyó la información acerca del referente y del navegador que utilizaba el visitante (véase la Figura 1).

Ante esta información, la acción más común era parsear la línea que generaba el servidor en forma automática en información con el fin de buscar los errores que se generaban en la recuperación de los archivos de la página web (el consabido error 404, "Not found", mayoritariamente). Hasta que aparecieron las primeras herramientas de análisis web, se reconoce a GetStats (predecesora de wwwstats) como una de las primeras herramientas que separaba en términos los diferentes datos que integraban un log de aquel entonces y lo ofrecía como información legible por humanos.

Claro que para que estas herramientas llegaran más allá de los informáticos (y les interesaran a los gestores de *marketing*, entre otro público), tenía que popularizarse el uso de la web y a su vez comenzar la era del comercio web, ya que una vez que se produce este cambio en la red, con el auge de las .com, es cuando realmente empieza a cobrar importancia la forma en que los usuarios navegan la web y comienzan a estudiarse sus patrones de visita. Es por ello que los nuevos analistas necesitaban segmentar los *logs* de los servidores de forma avanzada, y así nacen las soluciones que funcionan a través de la inclusión de código (JavaScript) en las páginas del sitio web.

El análisis de *logs* del servidor presenta las ventajas:

- Privacidad de los datos: el software instalado reside en el servidor de la institución y de esta forma se asegura la imposibilidad de que estos datos sean utilizados con fines comerciales o de análisis por otra institución o persona que la que tiene acceso al servidor o al software instalado en él y sus reportes; es decir, sus administradores.
- Funciona también para sitios publicados únicamente vía LAN: para instalar otro tipo de soluciones es necesario contar con una conexión a internet, los programas que parsean la información del *log* del servidor local no necesitan estar conectados a internet para realizar ninguna operación y pueden recolectar información de sitios que no son publicados en forma abierta a la WWW.
- Funciona sin necesidad de ejecutar JavaScript: estas soluciones no necesitan ejecutar código de JavaScript, ya que utilizan como fuente de datos el *log* del servidor. De esta forma, los dispositivos que no ejecutan JavaScript (como algunos celulares), con otro tipo de softwares no son incluidos en los reportes, en cambio analizando el *log* del servidor lo son.
- Las páginas se cargan más rápido que con otras herramientas: Si se agrega código a las páginas web se acrecienta el tiempo de carga de las páginas, con estos mecanismos no se genera ningún tipo de incremento.

Como desventajas pueden mencionarse:

- Son más difíciles de configurar e implementar, y todo el trabajo recae en el departamento de Sistemas.

» El análisis de *logs* es mucho más trabajoso.
» No recupera tantos datos sobre las características del software que posee el usuario.

1.2. Otros mecanismos de monitoreo web

Solamente por mencionar algunas de las otras técnicas de recolección de datos vía web:

» Contadores de visitas.
» Detección de paquetes.

1.3.- Inclusión de código en las páginas del sitio web (Page tagging)

Otra forma de obtener datos acerca del uso de los sitios web es a través de la inclusión de código JavaScript en las páginas del sitio web de las cuales quiere obtenerse información. Este código JavaScript lo brinda el software que se eligió como herramienta de analítica web y debe insertarse en cada página que quiere ser rastreada, o en algún lugar común que utilicen todos, las páginas web de determinado sitio para facilitar la tarea.

Este enfoque tiene como ventajas:

» Es muy fácil de implementar: una vez que se genera una cuenta de alguno de estos sistemas de analítica web el software brinda el código que hay que insertar en alguna parte de la etiqueta <head> de los archivos que componen el sitio web y el programa comienza a recolectar los datos de los visitantes.
» Se accede a los datos desde cualquier sitio, no hace falta estar logueado en el servidor donde se encuentra alojado el sitio web.
» No hace falta tener acceso físico al servidor donde se encuentra alojada la página web, sí hay que tener permisos de escritura de los archivos que conforman el sitio web.
» La mayor parte de los desarrollos están orientados a esta área, lo cual posibilita tener acceso a recursos actualizados en forma continua sobre analítica web.
» Presenta mayor cantidad de datos acerca del usuario, tales como resolución de pantalla, seguimiento de los eventos y títulos de las páginas web que visita.
» Permite leer las *cookies* del usuario, con lo cual puede acceder a la información de si es un usuario que está logueado en el sistema o si no lo está.
» Al mismo tiempo, presenta las desventajas:
» Lee solamente información de las páginas que se cargan (no registra información de errores 404).
» Requiere un intérprete de JavaScript en el cliente, los cuales no están presentes en los celulares, por ejemplo. Además, los usuarios que deshabilitan las *cookies* se contabilizarán como una visita única cada vez que ingresen al sitio.

1.3.1.- Herramientas de analítica web

Las herramientas de analítica web registran las diferentes acciones que llevan a cabo los visitantes en su interacción con un sitio web, y cuando esto ocurre, la herramienta de analítica web tiene acceso a valiosa información que debe ser utilizada de acuerdo con las implicancias éticas y legales que toda utilización de datos sensibles y que permiten identificar a las personas requiere.

Si bien como ya establecimos todos los sitios web tienen un *log* en el servidor que recolecta información cada vez que un usuario hace una petición al mismo, el programa

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)

seleccionado para compilar los datos de analítica web tiene acceso a su dirección IP, nombre de usuario logueado si se trata del caso, horario en que accede, etc.

1.3.1.1. Piwik

Piwik es una herramienta de analítica web que utiliza page tagging para registrar la interacción del usuario con el servidor. Pertenece también al grupo de los FOSS, se encuentra desarrollada en Perl y utiliza MYSQL como motor de base de datos (Miller, 2012). Desarrollada como tecnología cliente / servidor, puede accederse al Piwik utilizando un navegador web[4].

4. Para obtener más datos acerca de Piwik puede accederse al sitio oficial: http://piwik.org/

Para entender la relevancia que adquiere Piwik dentro del mercado de herramientas de analítica web, cabe mencionar que Avinash Kaushik; quien es una referencia obligada dentro de la analítica web, teniendo en su currículum vitae inclusive haber realizado consultorías de analítica web para Google, es proclive a su uso. Kaushik (2010), siendo uno de los autores más influyentes en estos aspectos, publicó: "si usted tiene una empresa orientada a lo técnico, no confíe en [los servicios de web analitics que proveen] Google o Yahoo! y descubra la aventura, que yo recomiendo personalmente, de usar Piwik. Una solución fantástica, se ha actualizado constantemente en estos dos años que la seguí" (traducción del autor, http://www.kaushik.net/avinash/best-web-analytics-tools-quantitative-qualitative/). Además, para aquellas bibliotecas que poseen servidores dedicados a un SIGB funcionando, no es difícil implementar una herramienta de web analytics tal como Piwik, ya que este sistema corre bajo MYSQL, utilizando algunas extensiones de PHP para su consulta on-line.

Algunos de los puntos sobresalientes de Piwik son:

» Software Open Source.
» Frecuencia de actualización.
» Facilidad de instalación (solamente tiene que copiarse una carpeta al servidor, generar una base de datos en MYSQL y correr el instalador vía http).
» Los datos son preservados por la institución.
» Los datos pueden ser anonimizados.
» Gran cantidad de sitios y foros de soporte técnico.
» Capacidad de configurar variables a posteriori.

1.3.1.2. Google Analytics

Por otro lado, existe también el motor de analítica web ofrecido por la empresa Google. Este motor tiene como principal ventaja que no necesita ser instalado para funcionar, ya que el servicio es gratuito, en la nube de Google. A su vez, la principal desventaja que se encuentra es justamente esta "ventaja", ya que si bien no necesita instalarse por funcionar bajo la nube de Google, esto como contrapartida no garantiza la privacidad de los datos de los usuarios (indispensable en una biblioteca). Los puntos sobresalientes de Google son:

» Software privativo (no se conoce el código fuente).
» Frecuencia de actualización.
» No necesita instalarse.
» Los datos son preservados por Google.
» Los datos pueden ser anonimizados (al menos para el cliente, no pudiendo verificarse la anonimización completa de los datos).
» Extensísima cantidad de sitios y foros de soporte técnico.
» Incapacidad de configurar variables a posteriori.

2. Comparativa entre Piwik y Google Analytics desde la privacidad

Por lo antedicho, podemos observar que Google Analytics es una herramienta muy potente y realmente fácil de configurar (no requiere instalación propiamente dicha), aunque uno de los aspectos que no debe ser descuidado en una biblioteca es la privacidad, ya que Google Analytics presenta una política de privacidad que, como todas las políticas de privacidad de Google, se reserva el derecho a modificar cuando haya un cambio en el servicio[5] o en las leyes del país (Estados Unidos) de la empresa que brinda el servicio. Este no es un detalle menor, ya que como se recordará es en Estados Unidos donde la privacidad de los ciudadanos suele ser atacada hasta en las bibliotecas (Elliott, 2013).

5. Google Analytics Terms of Service. Disponible en: http://www.google.com/analytics/terms/us.html

Es teniendo en cuenta estos aspectos que la Bibliotecología debería plantearse si es una buena decisión implementar Google Analytics como herramienta de analítica web, ya que en caso de implementarlo en un catálogo en línea de biblioteca podría recuperarse cuáles son las cadenas de búsqueda, los registros sobre los que el usuario hizo click para solicitar mayor información bibliográfica, y qué registros reservó, envió por mail o imprimió para su futura consulta. En este caso, siendo Google la empresa que se seleccionó para manejar la analítica web, probablemente el usuario acceda desde una IP determinada, que sea la misma que utiliza para manejar su hipotética cuenta de Google (Gmail, GoogleDocs, GoogleDrive o Play, etc.), o esté logueado en el sistema de biblioteca, con sus credenciales unívocas de usuario, con lo cual la identificación de dicho usuario en una empresa cuya capacidad para acceder a datos es enorme sería sencillamente muy fácil.

El modelo que se sugiere en este trabajo, teniendo en cuenta estos aspectos y la posibilidad de acceder a herramientas recomendadas por trabajadores de Google, es utilizar Piwik en contrapartida, ya que ofrece la posibilidad de alojar los datos en forma local y que únicamente los administradores o las personas determinadas a tal efecto puedan acceder a los mismos.

Conclusiones y recomendación final

Existe un creciente interés en implementar herramientas de analítica web en las bibliotecas, el cual puede explicarse desde paradigmas comerciales o de mejora de servicios a los usuarios, y se reconoce la necesidad de evaluar los servicios con información de calidad, fiable y recabada por instrumentos objetivos. A pesar de ello, las bibliotecas tienen un compromiso con el usuario y la gestión y uso de sus datos personales a los cuales tiene acceso cada vez que accede a un sitio web perteneciente a la institución.

Teniendo en cuenta las características funcionales, dos herramientas de analítica web fueron evaluadas desde el punto de vista de la privacidad: Google Analytics y Piwik, determinando que Piwik era la única que cumplía con ciertos requisitos mínimos desde la perspectiva de potencial uso de datos personales de los usuarios, aspecto no menor de la cuestión en una biblioteca en un mundo cambiante y ávido de bases de datos e información acerca de las preferencias de los navegantes web.

Otras investigaciones deberían profundizar en la razón que fundamenta que una institución (American Library Association) que estuvo en contra del Acta Patriótica de Estados Unidos hoy edita una publicación (*Library Technology Reports*) que recomienda la utilización de Google Analytics como herramienta de analítica web.

Referencias Bibliográficas

» Croll, Alistair y Sean Power. 2009. *Complete Web Monitoring*. Beijing: O'Reilly.

» Elliott, Justin. 2013. *Remember When the Patriot Act Debate Was All About Library Records?* <http://www.propublica.org/article/remember-when-the-patriot-act-debate-was-about-library-records> [Consulta: 15 Julio 2015].

» Google Analytics Terms of Service. <http://www.google.com/analytics/terms/us.html> [Consulta: 15 Julio 2015].

» Kaushik, Avinash. 2010. *Best Web Analytics 2.0 Tools: Quantitative, Qualitative, Life Saving!* <http://www.kaushik.net/avinash/best-web-analytics-tools-quantitative-qualitative/> [Consulta: 15 Julio 2015].

» Marek, Kate. 2011. Web Analytics Overview. En *Library Technology Reports. Using Web Analytics in the Library.* Vol. 47, No. 5, 5-10. <https://journals.ala.org/ltr/article/view/4233> [Consulta: 15 Julio 2015].

» Miller, Stephan A. 2012. *Piwik Web Analytics Essentials: A complete guide to tracking visitors on your websites, e-commerce shopping carts, and apps using Piwik Web Analytics*. Birmingham: Packt Publishing.

» Piwik. <http://piwik.org/> [Consulta: 15 Julio 2015].