DOI: http://dx.doi.org/10.11588/ip.2016.2.32678

Forschungsdaten: http://dx.doi.org/10.5281/zenodo.60515

Christian HAUSCHKE[1]

Third-Party-Elemente in deutschen Bibliothekswebseiten

Zusammenfassung

Einbindung von Third Party Elements (TPE) in Webseiten erlaubt es Dritten, webseitenübergreifend Nutzer zu identifizieren und ihr Informationsverhalten zu speichern. 4753 Bibliothekswebseiten wurden im Rahmen dieser Untersuchung mit der Open-Source-Software webXray auf die Einbindung von TPE untersucht. 54,77 % der analysierten Webseiten wiesen solche TPE auf. 18,94 % setzten Cookies ein, 44,81 % banden Javascript von Drittanbietern ein. Google-Services dominieren die TPE-Anbieterliste, sie werden in 30,02 % der untersuchten Webseiten verwendet.

Schlüsselwörter

Datenschutz; Bibliothekswebseite; Third-Party-Elemente

Third Party Elements in German library websites

Abstract

Embedding of Third Party Elements (TPE) in websites allows third parties cross-site identification of users and evaluation of their information behavior. 4753 library websites were examined with the open source software webXray regarding their use of elements, which can be used to inform third parties, if a web page was retrieved by a browser. 54.77% of the analyzed sites reported such TPE. 18.94% used third party cookies, 44.81% used third party Javascript. Google services dominate the TPE provider list, they are used in 30.02% of the investigated sites.

Keywords

Privacy; Library website; Third party elements

Erklärung

Christian Hauschke ist Mitglied der Redaktion und des Editorial Boards von Informationspraxis.

1 Bibliothek der Hochschule Hannover

Inhaltsverzeichnis

Einleitung

In den ersten Jahren des World Wide Webs lagen Webseiten in vollem Umfang auf einem einzigen Server, von dem Clients sie abrufen konnten. Dies ist schon seit einigen Jahren nicht mehr die Regel. Mehr und mehr werden Third-Party-Elemente (TPE) in Webseiten eingebunden, der Inhalte oder Funktionalitäten von weiteren, ursprünglich nicht vom User abgerufenen Servern hinzufügt. TPE können Video-Clips sein, Präsentationsfolien oder Bilder, die auf Servern von Drittanbietern liegen.

Bei jeder dieser Anfragen wird ein so genannter http-Request abgesetzt, der Informationen über den User liefert.

> IP: 123.45.67.89
> DATE: [19/August/2016:20:15:56 +0000]
> REQUEST: “GET /zaehlpixel.png HTTP/1.1”
> REFERER: http://bibliothekswebseite.de/datenschutz.html
> USER-AGENT: “Mozilla/5.0 (Windows NT 6.1; WOW64; rv:47.0) Gecko/20100101 Firefox/47.0”

In diesem Beispiel wird von einem User mit der IP-Adresse 123.45.67.89 eine Webseite „http://bibliothekswebseite.de/datenschutz.html“ aufgerufen, in der das TPE zaehlpixel.png enthalten ist. Dadurch erfährt der TPE-Lieferant die IP-Adresse, welcher Browser (User-Agent) eingesetzt wurde und über den Referer auch, welche Webseite der User abgerufen hat. Dieses Beispiel ist (Libert, 2015a, S. 3546) entliehen, dort finden sich auch weiterführende Informationen über Webtracking und die zugrundeliegenden Mechanismen und Methoden. Wie die übermittelten Informationen aussehen können, kann man u.a. mit Rex Swain’s HTTP Viewer (Swain, 1999) ausprobieren.

Die Nutzung von TPE hat stark zugenommen. Vor allem, da sie Webseiten-Entwickler auf vielfältige Art unterstützen. Zu den wichtigsten TPE gehören Webtracking-Tools wie zum Beispiel Google Analytics. Diese Werkzeuge erlauben es Webseiten-Inhabern, die Nutzung ihres Angebots zu evaluieren und es darauf aufbauend zu verbessern. Durch die weite Verbreitung dieser Tools können die Anbieter (wie z.B. Google) umfassende Benutzerprofile

erstellen, indem sie einen Nutzer über verschiedene Webseiten verfolgen. Aus der Nutzung bestimmter Webseiten – zum Beispiel Ratgeber, medizinische Einrichtungen oder Selbsthilfegruppen – lassen sich Schlüsse über den Gesundheitszustand einzelner Personen ziehen.

Cookies sind Dateien, in denen Webseitenbetreiber Informationen zur Identifikation oder auch webseitenspezifische Einstellungen lokal auf dem Rechner der Nutzer hinterlegen können. Cookies von Drittanbietern können beispielsweise durch Social-Media-Buttons zum Teilen von Inhalten angelegt werden.

Ein anderes Beispiel für TPE sind Javascript-Bibliotheken, die Entwicklern die zügige Integration komplexer Funktionen in eine Webseite erlauben. Durch sie kommen auch Javascript-Laien schnell zu erstaunlichen Ergebnissen. (Brehm, 2016) beschreibt Mechanismen (sogenannte Fingerprinting-Techniken) mit deren Hilfe Werbetreibende mit diesen und weiteren Hilfsmitteln webseitenübergreifend Nutzer identifizieren können. Sogar der Ladezustand von Smartphone-Akkus könne laut Brehm zu Hilfe genommen werden. Weitere technische und juristische Implikationen beschreiben (Wambach, Schulte und Knorr, 2016).

Demgegenüber stehen öffentliche und wissenschaftliche Bibliotheken, die sich laut Code of Ethics verpflichtet fühlen, die Privatsphäre ihrer Nutzerinnen und Nutzer – zum Kundenbegriff siehe (Zschau und Jobmann, 2013, S. 6ff.) – zu respektieren:

> "Wir respektieren die Privatsphäre unserer Kundinnen und Kunden. Wir speichern personenbezogene Daten nur zur Erbringung unserer Dienstleistung und nur im gesetzlichen Rahmen. Anderen Behörden stellen wir Benutzerdaten nur im engen Rahmen der gesetzlichen Vorschriften zur Verfügung." (Bibliothek und Information Deutschland, 2007)

Auch wenn hier nur von der Speicherung durch die Bibliothek und die Weitergabe an Behörden, also staatliche Einrichtungen, gesprochen wird, sind die Daten der Bibliotheksnutzerinnen und –nutzer ebenso schutzbedürftig gegenüber anderen, nichtstaatlichen Stellen, wie z.B. Firmen. Im Folgenden soll untersucht werden, ob die Webseiten deutscher Bibliotheken durch TPE die Sammlung von Informationen über ihre Nutzerschaft ermöglichen.

1 Methoden

Lobid.org stellt Informationen über bibliothekarische Einrichtungen als Linked Data über eine Schnittstelle (die lobid-API) zur Verfügung. Es stellt eine einfache Lösung dar, schnell Erhebungen über eine große Zahl von Bibliotheken zu tätigen. Dank der Hilfe von Fabian Steeg konnte mittels zweier Linux-Shell-Kommandos in wenigen Sekunden eine Liste mit URLs von Webseiten deutscher Bibliotheken zusammengestellt werden.

Mit dieser Methode wurden am 5. Mai 2016 insgesamt 5292 Adressen gesammelt, bei denen in wenigen Fällen die URL korrigiert werden musste. Abgesehen von wenigen offensichtlichen Fehlern (z.B. „http//:“ statt „http://“) wurden keine Korrekturen am Datensatz vorgenommen. Die verwendeten URLs und die zur Sammlung verwendete Methode sind unter (Steeg und Hauschke, 5. Mai. 2016) zu finden.

Harvesting und Auswertung fanden mit der Open-Source-Software webXray (Libert und Rodriguez, 2016) statt, mit der Libert unter anderem untersuchte, welche Drittparteien über den Abruf von gesundheitsbezogenen Webseiten informiert werden (s. Libert, 2015b). webXray ruft Webseiten auf und stellt für jede Seite fest, welche Server beim Ansurfen der Webseite kontaktiert werden.

Nach der automatischen Entfernung von Dubletten durch webXray – die Top Level Domain (TLD) sub.uni-goettingen.de tauchte zum Beispiel vier Mal auf –verblieben 4866 Adressen, die von webXray in der Zeit vom 7. bis zum 11. Mai eingelesen wurden.

Einschränkungen:

1. Nicht alle Adressen sind eindeutig einer Bibliothek zuzuordnen. http://polizei.sachsen.de führt zum Beispiel zu einer Webseite der sächsischen Polizei; zahlreiche, gerade kleinere öffentliche Bibliotheken verweisen auf die Webseiten ihrer kommunalen Träger.
2. Es wurde nur die Startseite durchsucht. Wenn TPE auf Unterseiten, im Bibliothekskatalog oder im Open-Access-Server verwendet werden, ist das im Rahmen dieser Untersuchung unsichtbar.

Nach dem Auslesen liefert webXray verschiedene automatisch generierte Statistiken. Insgesamt analysierte webXray 4753 Seiten (97,6 %), weitere 113 Seiten (2,3 %) konnten nicht untersucht werden.

2 Ergebnisse

Die Analyse der vorliegenden URLs haben verschiedene Ergebnisse erbracht. Schon vor der Auswertung durch webXray ist zu erkennen, dass nur 173 der ursprünglich eingelesenen URLs eine Webadresse mit https angeben, also verschlüsselt sind. Hier ist zu berücksichtigen, dass durchaus Redirects auf https-Adressen vorliegen könnten. Dies wurde bei 20 zufälligen Stichproben überprüft. Dabei konnte kein Redirect beobachtet werden.

Cookies von Drittanbietern wurden von ca. 18 % der untersuchten Seiten eingesetzt (Tabelle 1).

Cookies insgesamt	3354
Seiten mit Cookies	900
Seiten mit Cookies (%)	18

Tabelle 1 - Cookies in Bibliothekswebseiten

Insgesamt wurden von den untersuchten Webseiten 32.687 TPE angefordert, davon wurden 91 % erfolgreich übermittelt. Knapp über die Hälfte der Webseiten (54 %) greifen auf TPE zurück.

Insgesamt waren TPE in 2603 (54 %) der untersuchten Webseiten enthalten. 32687 TPE wurden angefordert, 30036 (91 %) erfolgreich übermittelt. Google führt die Liste der eingebundenen Drittelemente mit deutlichem Vorsprung an. Fast ein Drittel aller analysierten Seiten rufen ein oder mehrere Elemente von Google-Servern ab. Erst mit sehr großem Abstand folgen Face-

book, jQuery Foundation oder Monotype Imaging. Alle 18 Organisationen mit mehr als 10 betroffenen Webseiten sind der Tabelle 2 zu entnehmen.

Rang	Organisation	Herkunft	Betroffene Webseiten	Anteil in %
1	Google	US	1427	30,02
2	Facebook	US	129	2,71
3	jQuery Foundation	US	114	2,40
4	Monotype Imaging	US	84	1,77
5	MaxCDN	US	65	1,37
6	Twitter	US	64	1,35
7	ReadSpeaker	NL	51	1,07
8	Oracle	US	49	1,03
9	Amazon	US	48	1,01
10	Adobe	US	44	0,93
11	Cloudflare	US	40	0,84
12	Automattic	US	24	0,50
13	Yahoo!	US	19	0,40
14	comScore	US	14	0,29
15	New Relic	US	13	0,27
16	Akamai	US	12	0,25
17	Media Math	US	12	0,25
18	Rubicon Project	US	11	0,23

Tabelle 2 - TPE nach Organisation

Google führt die Rangliste sehr deutlich an und ist in fast einem Drittel der untersuchten Webseiten vertreten. Weitere Dienste sind weitaus seltener zu finden.

Die am häufigsten verwendeten TPE nach TLDs (Tabelle 3) weichen in einigen Punkten von den TPE nach Organisationen ab, da einige Organisationen verschiedene TPE-Domains anbieten. Google hat verschiedene Domains (googleapis.com oder youtube.com), auch Facebook bietet verschiedene Dienste an.Nun rücken auch Seiten wie Kirchenserver.info und verwaltungsportal.de in die 20 verbreitetsten TPE-Lieferanten. Google bleibt insgesamt mit weitem Abstand am häufigsten eingebunden.

Rang	Top Level Domain	Herkunft	Betroffene Webseiten	Anteil in %
1	googleapis.com	US	939	19.76
2	gstatic.com	US	663	13.95
3	google-analytics.com	US	616	12.96
4	google.com	US	335	7.05
5	doubleclick.net	US	244	5.13
6	etracker.de	DE	138	2.90
7	etracker.com	DE	129	2.71
8	facebook.com	US	124	2.61
9	jquery.com	US	114	2.40
10	ytimg.com	US	95	2.00
11	fbcdn.net	US	94	1.98
12	youtube.com	US	86	1.81
13	facebook.net	US	83	1.75
14	bootstrapcdn.com	US	65	1.37
15	fonts.net	US	62	1.30
16	kirchenserver.info	DE	58	1.22
17	verwaltungsportal.de	Unknown	58	1.22

18	hbz-nrw.de	Unknown	56	1.18
19	twitter.com	US	55	1.16
20	readspeaker.com	NL	51	1.07

Tabelle 3 - TPE nach Top Level Domain

3 Fazit

Selbst wenn sich der Code of Ethics nur auf den Schutz von Nutzerdaten vor behördlicher Einsicht bezieht, darf die Achtung der informationellen Selbstbestimmung und der Schutz der Nutzer und Nutzerinnen an dieser Stelle nicht aufhören. Dies gilt umso mehr, wenn eine webseitenübergreifende Identifikation von Geräten – und damit in sehr vielen Fällen auch von Personen – durch Dritte durch Nichtverwendung von TPE verhindert werden kann. Bibliotheken können durch einfache Maßnahmen viel zum Schutz ihrer NutzerInnen beitragen. Der erst Schritt ist der Verzicht auf die verbreitetsten TPE, und dies sind nach Tabelle 3 Javascript-Bibliotheken (z.B. googleapis.com) und Webfonts.

- Verzicht auf Fremdhosting von Javascript-Bibliotheken. Sie können – statt auf den Servern von Drittanbietern - auch auf dem eigenen Server gehosted werden.
- Verzicht auf die Einbindung von Analyse-Diensten aus dritter Hand. Schon (Jürgens, Mandl und Womser-Hacker, 2010) haben auf die Möglichkeit hingewiesen, aus Datenschutzgründen selbstgehostete Open-Source-Lösungen wie z.B. PIWIK (http://piwik.org/) als Ersatz zu verwenden.
- Verzicht auf die Einbettung von Webfonts als TPE. Die Omnipräsenz von Webfonts gerade in populären Content Management Systemen (CMS) und dazugehörigen Themes hat dazu geführt, dass für einzelne CMS schon Erweiterungen verfügbar sind, die diesen Schritt erleichtern. Als Beispiel sei „Remove Google Fonts References“ (https://de.wordpress.org/plugins/remove-google-fonts-references/) für Wordpress genannt.

Was (Wambach, Schulte und Knorr, 2016) über den Gesundheitsbereich schreiben, lässt sich nahtlos auf das Bibliothekswesen übertragen:

> „Es ist dringend erforderlich, dass man sich vor allem im Gesundheitsbereich vergegenwärtigt, dass die eigene Webseitengestaltung die Bildung von Nutzerprofilen durch Dritte ermöglichen kann. Will man seinen datenschutz-rechtlichen Pflichten und dem Interesse der Nutzer, sich unverfolgt und entsprechend unbefangen zu informieren, gerecht werden, ist es geboten, den eigenen Webauftritt auf die Einbettung von Drittparteien hin zu überprüfen.“

Dieser Artikel soll einen Beitrag dazu leisten, sich der Verpflichtungen des Bibliothekswesens hinsichtlich der informationellen Selbstbestimmung ihrer NutzerInnen bewusst zu machen. Zur freien Verfügbarkeit von Informationen gehört immer auch die Möglichkeit, diese unbeobachtet nutzen zu können.

Weiterer Forschungsbedarf besteht hinsichtlich der Verbreitung von TPE in Katalogen, Datenbanken, Open-Access-Repositorien und anderen von Bibliotheken lizenzierten oder selbst angebotenen Diensten.

Quellen

Alle in diesem Quellenverzeichnis aufgeführten Webseiten wurden letztmalig am 19.08.2016 aufgerufen.

Bibliothek und Information Deutschland (2007): *Ethische Grundsätze der Bibliotheks- und Informationsberufe* [online] [Zugriff am: 28. Mai 2016]. URL: http://www.bibliotheksportal.de/themen/beruf/berufsethik/code-of-ethics-bid-2007.html

Brehm, Joachim (2016): Verräterische Merkmale - Neue Browser-Fingerprinting-Techniken - und wie man sich schützt. *c't*, (11), 144-145.

Jürgens, Julia, Mandl, Thomas, Womser-Hacker, Christa (2010): Das Potenzial von Web Analytics für Usability-Evaluierungen. In: A. Schmidt und J. Ziegler, Hg. *Mensch & Computer 2010. 10. fachübergreifende Konferenz für interaktive und kooperative Medien. Interaktive Kulturen.* Berlin, Boston: Oldenbourg Wissenschaftsverlag, S. 261-270. ISBN 978-3-486-85348-3.

Libert, Tim; Rodriguez, Dani (2016): webXray [Software]. First Release. URL: http://dx.doi.org/10.5281/zenodo.57272

Libert, Timothy (2015a): Exposing the Invisible Web: An Analysis of Third-Party HTTP Requests on 1 Million Websites [online]. *International Journal of Communication*, 9, 18. URL: http://ijoc.org/index.php/ijoc/article/download/3646/1503

Libert, Timothy (2015b): Privacy implications of health information seeking on the web [online]. *Communications of the ACM*, 58(3), 68-77. URL: http://dx.doi.org/10.1145/2658983

Steeg, Fabian; Hauschke, Christian (05.05.2016): URLs von Webseiten mit Typ Bibliothek aus Lobid.org. URL: http://dx.doi.org/10.5281/zenodo.50969. URL: http://dx.doi.org/10.5281/zenodo.50969

Swain, Rex (1999): *Rex Swain's HTTP Viewer* [online]. *See exactly what an HTTP request returns to your browser.* 20 Juli 2015. URL: http://www.rexswain.com/httpview.html

Wambach, Tim, Schulte, Laura, Knorr, Konstantin (2016): Einbettung von Drittinhalten im Web [online]. *Datenschutz und Datensicherheit - DuD*, 40(8), 523-527. URL: http://dx.doi.org/10.1007/s11623-016-0650-6

Zschau, Gerhard; Jobmann, Peter (2013): *Auf dem Weg zur demokratischen Bibliothek.* Berlin, Freie Universität. URL: http://demokratische-bibliothek.de/wp-content/uploads/2014/06/Masterarbeit_Jobmann_Zschau_DemPaed_.pdf

Autor

Christian HAUSCHKE
Bibliothek der Hochschule Hannover
Ricklinger Stadtweg 118, D-30459 Hannover
http://www.hs-hannover.de/bibl
christian.hauschke@hs-hannover.de