Received 19 February 2018
Revised 16 June 2018
Accepted 13 July 2018

Protecting privacy on the web

A study of HTTPS and Google Analytics implementation in academic library websites

Patrick O'Brien, Scott W.H. Young and Kenning Arlitsch
Montana State University, Bozeman, Montana, USA, and
Karl Benedict
University of New Mexico, Albuquerque, New Mexico, USA

Abstract

Purpose – The purpose of this paper is to examine the extent to which HTTPS encryption and Google Analytics services have been implemented on academic library websites, and discuss the privacy implications of free services that introduce web tracking of users.
Design/methodology/approach – The home pages of 279 academic libraries were analyzed for the presence of HTTPS, Google Analytics services and privacy-protection features.
Findings – Results indicate that HTTPS implementation on library websites is not widespread, and many libraries continue to offer non-secured connections without an automatically enforced redirect to a secure connection. Furthermore, a large majority of library websites included in the study have implemented Google Analytics and/or Google Tag Manager, yet only very few connect securely to Google via HTTPS or have implemented Google Analytics IP anonymization.
Practical implications – Librarians are encouraged to increase awareness of this issue and take concerted and coherent action across five interrelated areas: implementing secure web protocols (HTTPS), user education, privacy policies, informed consent and risk/benefit analyses.
Originality/value – Third-party tracking of users is prevalent across the web, and yet few studies demonstrate its extent and consequences for academic library websites.

Keywords Web analytics, HTTPS, Third-party tracking, Web privacy
Paper type Research paper

Introduction

Third-party tracking can occur when web analytics services, such as Google Analytics, are utilized to measure visitation to websites. These services provide information about website use and user behavior, which can help libraries improve their online services. However, the analytics services operate sophisticated mechanisms through extensive networks to track users and their behavior across sites, acquiring user demographics and behavioral patterns. The detailed tracking enabled by Google Analytics is often performed without the fully informed consent of individual users of the website. The extent to which Google Analytics services have been implemented within the domain of library websites has been unknown prior to this study. Unknown, also, has been the extent to which available privacy-protecting features have been implemented on those websites.

The library profession has long supported the principles of privacy, but tracking used by analytics service providers has rendered those principles nearly untenable. For example, without proactive efforts to mitigate their impact, browser cookies set by Google Analytics act as beacons for collecting and sharing user data through a vast network of commercial trackers. By understanding the extent and significance of web tracking and the available

Online Information Review
Vol. 42 No. 6, 2018
pp. 734-751
Emerald Publishing Limited
1468-4527
DOI 10.1108/OIR-02-2018-0056

privacy-protection mechanisms, libraries can begin to minimize their participation in third-party tracking on the web.

The results presented in this paper demonstrate conclusively that 279 academic libraries from around the world must do much more to ensure user privacy if they hope to maintain trust with their users. The principle of this trust is outlined in the privacy statements of the American Library Association (ALA), Coalition for Networked Information (CNI), National Information Standards Organization (NISO) and the International Federation of Library Associations and Institutions (IFLA).

In presenting our research, we first explain web tracking, web analytics and web privacy. We then detail our methods and results, followed by a discussion of the privacy implications of third-party web tracking. We conclude by offering recommendations for professional action and avenues for future research.

Literature review

Web tracking

The practice of third-party tracking on websites is widespread (Narayanan and Reisman, 2017), and has only increased in prevalence, variety and complexity over time (Lerner *et al.*, 2016; Englehardt and Narayanan, 2016). One of the most common trackers found on the Web is produced by the Google Analytics web service, which is used to measure the visitation to a website (Lerner *et al.*, 2016; Schelter and Kunegis, 2016). In exchange for this easy-to-implement and free-to-use analytics service, websites execute Google Analytics JavaScript code and pass user visit data to Google through browser cookies set by Google Analytics (Krishnamurthy and Wills, 2009). Such data are considered to be "leaked" if the user is unaware of its collection and does not consent to the data being shared with additional third parties (Sar and Al-Saggaf, 2013). An analysis of 1m websites found that nearly nine in ten websites leak user data to third parties without the user's knowledge (Libert, 2015).

The Google Analytics tracker is not designed to leak user data across sites on its own, but its tracking capabilities are enhanced when combined with Google AdSense, Google's popular cross-site advertising service that utilizes its Doubleclick tracker. When Google AdSense and Google Analytics have both been implemented in a website, the unique identifiers from each service can be linked by Google's Doubleclick tracker such that Google can create browsing profiles that track users across sites (Roesner *et al.*, 2012). Data leakage from Google Analytics can also occur when websites activate the additional Google tracking service known as Tag Manager, which allows for cross-site tracking and targeted advertising (Bashir *et al.*, 2016). Under these expanded tracking conditions, third-party trackers can match user behavior data with user profiles, thereby allowing users to be tracked and targeted across the web (Olejnik *et al.*, 2012; Falahrastegar *et al.*, 2016; Kalavri *et al.*, 2016). While data about Google Tag Manager and Google AdSense were collected during course of this study, full analysis is beyond the scope of this paper.

Data leakage and user profiling via web tracking represents a privacy issue for users because of a lack of transparency and the lack of opportunity for users to consent to the sharing of their tracked behavior. The following example illustrates this case:

> A user logs into Gmail and then visits a library website that has implemented Google Analytics or Google Tag Manager. This user then searches for tax relief resources through the library website. Because Google 1) identifies and authenticates users via their Google IDs and passwords and 2) identifies and authenticates the library website through Google Analytics or Tag Manager, Google can link users' library website activity to individual users' Google profiles. Depending on the library's Google implementation, this user activity may also be shared with Google's advertising network, which targets users with personalized ads, such as credit cards or personal loan services, even after the user has left the library web site.

This style of tracking is pervasive; Google was shown to be capable of tracking users on nearly 80 percent of the top 1m websites (Libert, 2015). Websites that implement Google Analytics and other Google tracking services are participating in the extensive network of third-party trackers that are capable of sharing user data across sites. It appears that in most cases, the user has neither knowingly or explicitly given informed consent for this type of data sharing, nor does the website owner fully understand the capabilities and consequences of web analytics and other third-party trackers. While this is a common practice when interacting with many sites on the web, academic libraries using Google trackers without proactively enabling user privacy features may have unwittingly violated the principles of user privacy expressed by the ALA, CNI, NISO and IFLA privacy guidelines.

Web privacy

The library science professional literature includes many contributions that detail the implementation, application and justification of Google Analytics for the purposes of web traffic analysis and service improvement (Hess, 2012; Barba *et al.*, 2013; Cohen and Thorpe, 2015; Fagan, 2014; Yang and Perrin, 2014; Conrad, 2015; Farney, 2016). User privacy is seldom mentioned in these articles and manuals. Yet, the user data collected by Google Analytics, such as search terms, user-agent software, geographical location,and time of day, can potentially be leaked to other third-parties via the network of web trackers. User privacy can be further undermined when third parties match behavior data with user profiles, thereby allowing users to be tracked and targeted across the web (Olejnik *et al.*, 2012). Certain Google Analytics implementation methods can help reduce its data collection capability and reduce library participation in cross-site user tracking. These mitigating techniques include IP anonymization[1], opt-out mechanisms[2] and secure HTTP connections. A secure HTTP connection, also referred to as HTTPS, can be activated with a secure digital certificate and proper configuration of the host server (Naylor *et al.*, 2014; Askey and Arlitsch, 2015). The use of HTTPS: ensures that communication over the public internet is encrypted; and when the certificate is provided by a trusted certificate authority, it provides a verification mechanism to assure users that the website they are visiting belongs to the domain name owner and server they have requested. Without HTTPS protection in place, user activities over wired or wireless networks can be observed and retained.

Best practices for search engine optimization indicate that websites should automatically redirect non-secure URL user requests (HTTP) to secure versions of the URL (HTTPS) by way of a permanent webserver redirect (Arlitsch and OBrien, 2013)[3]. These practices signal to users that site administrators are concerned with user privacy, thereby engendering trust.

Privacy has long been a concern of libraries (Million and Fisher, 1986; Garoogian, 1991; Johnston, 2000; Nichols Hess *et al.*, 2015), and defending privacy was much more attainable in the pre-digital world. Given the extent of third-party tracking on the internet, however, it is exceedingly difficult to implement analytics trackers like Google Analytics without compromising the privacy for users that libraries have championed. Library professional organizations have acknowledged the complexity of contemporary information privacy, and have modified privacy statements accordingly.

The ALA has published several statements and toolkits to help librarians achieve privacy for users. ALA describes its privacy and surveillance guidelines as an attempt "to balance the need to protect reader privacy with the needs of libraries to collect user data and provide personalized services[4]." In the Library Bill of Rights, the ALA offers a definition of privacy: "In a library (physical or virtual), the right to privacy is the right to open inquiry without having the subject of one's interest examined or scrutinized by others[5]." The ALA continues in the Library Bill of Rights: "Libraries should not share personally identifiable user information with third parties or with vendors that provide resources and library services unless the library has obtained the permission of the user or has entered into a legal

agreement with the vendor." ALA has offered additional calls-to-action through its Privacy Toolkit, which states, "For libraries to flourish as centers for uninhibited access to information, librarians must stand behind their users' right to privacy and freedom of inquiry[6]."

A recent Executive Roundtable Report of the Coalition for Networked Information (CNI) notes: "Libraries collecting data using Google Analytics are realizing they may be violating the ALA Library Bill of Rights[…] this is but one example of how easily convenient web-based service offerings can come with unexpected consequences[7]."

The NISO has released a document that outlines 12 privacy principles for third-party e-resource systems[8]. The IFLA Statement on Privacy in the Library Environment recommends: "Library and information services should reject electronic surveillance and any type of illegitimate monitoring or collection of users' personal data or information behavior that would compromise their privacy and affect their rights to seek, receive and impart information[9]." IFLA further identified that "the rapid advancement of technology has resulted in increasing privacy implications." From within this context of networked complexity and third-party tracking, the Library Freedom Project has drafted the First Library Digital Privacy Pledge, which aims to increase the implementation of HTTPS on library websites, and has gained 21 endorsements from membership organizations, public and academic libraries, and vendors, as of this writing.[10] See Table I for a summary of privacy statements from professional library organizations.

A survey of librarians' attitudes toward privacy found that 97 percent of respondents agree or strongly agree that libraries should never share personal information and circulation or internet records without authorization or a court order (Zimmer, 2014). In the same survey, 76 percent of respondents feel that libraries are doing all they can to prevent unauthorized access to individual's personal information and circulation records; however, a different survey investigating the configuration of public internet terminals showed that many libraries have not installed ad-blocking and privacy-protecting features on web browsers, nor do they offer instruction to users regarding web privacy

Organization	Statement title	Statement excerpt
American Library Association (ALA)	Library Bill of Rights – Interpretation of Privacy	In a library (physical or virtual), the right to privacy is the right to open inquiry without having the subject of one's interest examined or scrutinized by others
Coalition for Networked Information (CNI)	Privacy in the Age of Analytics	Libraries collecting data using Google Analytics are realizing they may be violating the ALA Library Bill of Rights…this is but one example of how easily convenient web-based service offerings can come with unexpected consequences
National Information Standards Organization (NISO)	NISO Privacy Principles	Libraries, publishers and software providers have a shared obligation to foster a digital environment that respects library users' privacy as they search, discover and use those resources and services
International Federation of Library Associations and Institutions (IFLA)	Privacy Statement	Library and information services should reject electronic surveillance and any type of illegitimate monitoring or collection of users' personal data or information behavior that would compromise their privacy and affect their rights to seek, receive and impart information
Library Freedom Project	Digital Privacy Pledge	Library services and resources should be delivered, whenever practical, over channels that are immune to eavesdropping

Table I.
Privacy statements – professional library organizations

(Gardner and Groover, 2015). As other authors neatly summarize, "many websites use [Google Analytics and other click-tracking mechanisms], and their utility within library systems is an ongoing debate as we balance the needs of reliable metrics with patron privacy" (Caro and Markman, 2016). The use of Google Analytics on library websites is ubiquitous, yet the tension between web analytics and web privacy demands further investigation to ensure that libraries are in fact doing all we can to prevent unauthorized and unwanted data sharing.

In response to the perceived lack of professional knowledge regarding the extent of third-party web tracking on library websites, we have conducted a privacy audit that empirically measures the extent of Google Analytics services and related privacy-protection features on library websites. While many librarians agree that libraries should not share user data, their use of Google Analytics services without implementing available privacy-protection features signal that libraries are not doing all they can to prevent user data leakage. Understanding the prevalence of tracking and privacy infrastructure is a fundamental first step for taking concerted professional action that will benefit the privacy of users. As Lerner *et al.* (2016, p. 997) assert:

> Measurement studies of web tracking are critical to provide transparency for users, technologists, policy-makers, and even those sites that include trackers, to help them understand how user data is collected and used, to enable informed decisions about privacy, and to incentivize companies to consider privacy.

To motivate the larger community of library professionals, we must first grasp the nature and extent of web tracking that occurs on library websites.

Research questions

RQ1. Do libraries implement HTTPS with proper redirect practices?

Does the library protect privacy with a secure connection (via HTTPS) between the user's browser and the library's website? Does the library use a permanent redirect to enforce the use of secure connections? Does the library redirect secure page requests to a non-secure version of the page in violation of recommended practice?

RQ2. Do libraries that use Google Analytics implement the available privacy-protection measures?

Does the library use Google Analytics? If the library is using Google Analytics, does it protect user privacy via a secure connection between the library website and Google's servers? If the library is using Google Analytics, does it obfuscate individual user tracking using Google's IP Anonymization feature?

Methodology

Webometrics is a subset of interrelated library and information science empirical research methodologies, whose relationship can be visualized as overlapping concentric circles beginning at the outer circle with informetrics and then moving inward toward bibliometrics, scientometrics, cybermetrics and webometrics (Björneborn and Ingwersen, 2004). As a family of methodologies, informetrics and its subsets comprise a relatively small percentage of the published library and information science research (Togia and Malliari, 2017).

Webometrics was originally proposed as a research methodology in the late 1990s when it became apparent that longstanding informetric and bibliometric methods could be applied to the content and structure of the World-Wide-Web (Almind and Ingwersen, 1997).

Webometrics initially focused on statistical analyses of word and phrase frequencies, citations, characteristics of authors and publications, and rankings and impact factors, but the definition of the new methodology evolved quickly. "Link structures and search engines" were added (Björneborn and Ingwersen, 2001), and the definition was then further expanded to include "quantitative aspects of the construction and use of information resources, structures and technologies on the Web" (Björneborn and Ingwersen, 2004). It is this expanded definition that guides the current research, which fundamentally investigates the prevalence of security and privacy structures in academic library websites.

Research methods

Within the webometrics methodology, our data-gathering method can be classified as covert observation research, a social sciences research technique used to observe participant behavior without revealing the identity or presence of the researcher (Punch, 2014; Taylor *et al.*, 2016). Covert observation has been used in many fields to gather both qualitative and quantitative data. In the health care field, for instance, it was used to discreetly observe behavior by participants in online communities that support or disparage eating disorders (Brotsky and Giles, 2007). It has also been used in quantitative marketing research, a business discipline that involves "collection of data gauging respondents' reactions to stimuli and perceptions of a product" (Kurian, 2013), but its emphasis in this discipline is on the agreement reached by the buyer and the seller. Covert observational techniques can raise ethical concerns when the subjects are people (Hallenberg, O'Neill, and Tong, 2015; Stanley and McLaren, 2007). However, the observed subjects in the current research are the information structures publicly hosted on machines; specifically, we observe the presence or lack of HTTPS and the presence or lack of the Google Tracking Code. As such, there are no ethical concerns associated with this research.

Research design

Our research examines the website home pages of 279 US and international academic libraries. The study population included libraries with one or more memberships in the following organizations as of March 4 and 5, 2016:

- Association of Research Libraries (ARL)[11];
- OCLC Research Library Partnership (OCLC-RLP)[12]; and
- Digital Library Federation (DLF)[13].

These organizations were selected due to: mission statements focused on "research" and "libraries"; and each organization has a membership exceeding 100 libraries[14]. The study population was audited on October 5, 2016 by requesting the publically-available HTML pages listed on each organization's membership page and logging the research library's webserver response. The study includes 448 unique URLs from 279 libraries in 16 countries. Geographically, the data set represents 344 unique URLs published by 211 US libraries, 30 URLs by 16 Canadian libraries, 33 URLs by 22 UK libraries and 41 URLs from 30 libraries in other countries. The full data set is available through Zenodo[15]. The process used to generate this data set, including scripts and documentation, is available for verification and replication as open source code through GitHub[16].

The following procedures are presented as step-by-step outlines, organized by research question.

RQ1: do libraries implement HTTPS with proper redirect practices?

In order to answer this research question, we completed three main steps of analysis for the websites in our study population.

Step 1. We determined whether the library offers a secure connection (HTTPS) between the user's browser and the library's website:

(1) Check for a digital certificate (HTTPS). This test was accomplished by requesting a secure connection to each URL.

- For example, www.unm.edu/libraries.html is a unique URL that can be requested with a non-secure (http://) or secure (https://) connection. For this test, the secure connection, www.unm.edu/libraries.html was requested and the sever response was logged.
- The test is "true" if the Library webserver resolved to a URL containing "https://."

Step 2. We determined whether the library has implemented a permanent redirect to enforce the use of secure connections, i.e. connecting via HTTPS.

(1) Check if the server redirects non-secure requests to secure connections of the page requested:

- For example, when the non-secure URL http://scholarworks.montana.edu/ is requested, does the webserver respond with an HTTP 301 permanent redirect message to the secure URL https://scholarworks.montana.edu/?
- This test is "true" if the library's webserver uses an HTTP 301 redirect of non-secure URL requests (http://) to secure (https://) versions of the URL.

Step 3. We determined whether the library redirects secure page requests to a non-secure version of the page.

(1) Check if the server redirects secure page request to a non-secure connection:

- For example, if a secure URL is requested (e.g. https://scholarworks.montana.edu/) does the webserver redirect the user to a non-secure version of the page (e.g. http://scholarworks.montana.edu)?
- This test is "true" if the library's webserver redirects the user to a non-secure page (http://) without first delivering an "HTTP 404 Page Not Found" message.

R2: do libraries that use Google Analytics implement the available privacy-protection measures?

In order to answer this research question, we completed three main steps of analysis for the websites in our study population.

Step 1. We determined whether the library website has implemented Google Analytics. As a corollary to this step, we also determined whether the library website has implemented Google Tag Manager. Below are the test requirements for this step:

(1) Each webpage was analyzed for the presence of the Google Analytics tracking code and, separately, for the Google Tag Manager tracking code, identified by unique markers that met the following criteria:

- for Google Analytics, the presence of a character string unique to and required by either Universal Analytics or Classic Analytics[17]; and
- for Google Tag Manager, the presence of a character string unique to and required by Google Tag Manager[18].

The use of Google Analytics was determined to be true if any of the library home pages contained the character strings unique to and required by Universal Analytics, Classic Analytics or Tag Manager.

Step 2. If a website tested positive for the tracking codes of either Google Analytics or Google Tag Manager, we then determined if the website was using available features to protect user privacy. First, we determined whether the library implemented a secure HTTPS connection between the library web server and Google's web server. To complete this step, we analyzed each website's source code for the presence of either "forceSSL"[19] or www.googletagmanager.com/[20].

Step 3. If the website is using Google Analytics, we determined whether it has implemented the Google Analytics IP anonymization feature to obfuscate user tracking. To complete this step, we analyzed each website's source code for the presence of "anonymizeIp[21]."

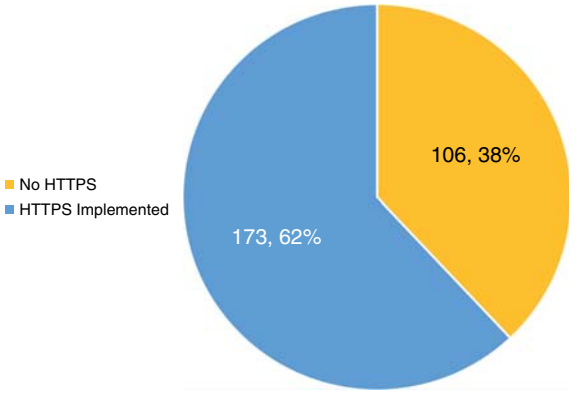
We tested for the presence of these two elements – forceSSL and anonymizeIp – because these are the only two Google Analytics privacy mechanisms that were observable at the time of this writing.

Results and discussion

Our study results indicate that libraries are not doing all they can to protect user privacy on the web. Analysis of the results of our testing follows.

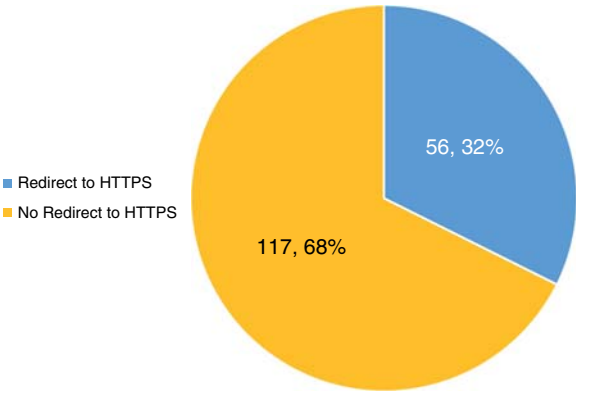
Results summary

- Of the libraries in our study population ($n = 279$), 173 (62 percent) have implemented basic encryption technology via HTTPS (see Figure 1).
- Of the libraries that have implemented HTTPS ($n = 173$), 56 (32 percent) implemented a permanent redirect from HTTP to HTTPS to ensure that HTTPS is used at all times when communicating with users (see Figure 2).
- Of the websites in our study population ($n = 279$), we also found that 43 (15 percent) have implemented a redirect in the inverse direction: from HTTPS to HTTP, thus allowing user activity to occur over a non-secure connection without informed consent – even when the user specifically requests a secure connection.
- Of the websites in our study population ($n = 279$), 245 (88 percent) have implemented either Google Analytics or Google Tag Manager (see Figure 3).
- Of the websites that implemented Google tracking code ($n = 245$), 3 (1 percent) had implemented HTTPS connections between the library's web server and Google's



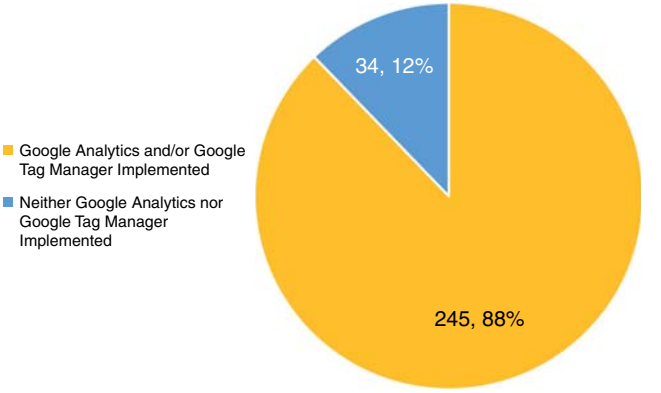
Note: $n = 279$

Figure 1. HTTPS implementation for academic library home pages



Note: $n=173$

Figure 2.
Redirect implementation for academic library websites with HTTPS



Note: $n=279$

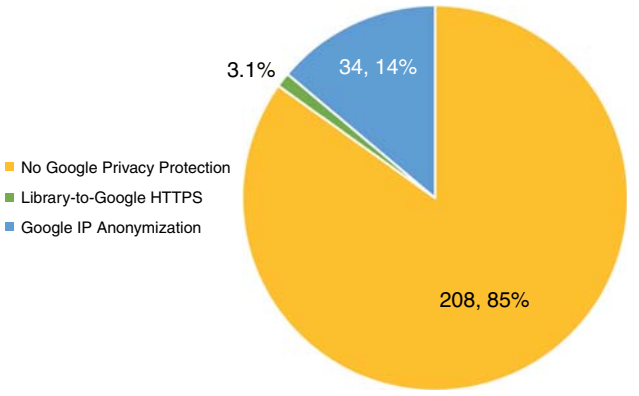
Figure 3.
Implementation of Google analytics and/or Google tag manager for academic library home pages

web servers, 34 (14 percent) had implemented Google's feature to obfuscate user identification via IP anonymization and 0 (0 percent) implemented both of these available privacy features (see Figure 4).

RQ1: do libraries implement HTTPS with proper redirect practices?

The first test determined whether the library has provided the opportunity for a secure HTTPS connection between the user's browser and the library's website. If so, did the library automatically redirect non-secure URL requests (i.e. http://) to a secure version of the URL (HTTPS) by way of a permanent webserver (HTTP 301) redirect? If a user requested a secure connection and one is not available, did the library inform the user that a secure connection is not available or did the library redirect the user to a non-secure connection without informed consent?

Our results indicate that HTTPS implementation on library websites is not widespread, and many libraries continue to offer non-secure connections without an automatically enforced redirect to a secure connection. In our study population, 62 percent had implemented HTTPS. Of those, only 32 percent automatically redirected non-secure



Note: $n = 245$

Figure 4.
Privacy-protection features for academic library home pages with Google analytics and/or Google tag manager

requests to secure requests. Furthermore, we found that 46 of the 106 websites without HTTPS available (43 percent) were redirecting secure requests made by users to non-secure URLs without notifying the user. This practice undermines privacy in a number of ways: first, by failing to inform the user that a secure HTTPS version of the page is not available; second, without obtaining informed consent before the user is offered non-secure web pages; and third, by increasing the risk of Man-In-The-Middle attacks of library website users connecting via non-secure Wi-Fi access points, such as coffee shops, or compromised wired network connections[22].

RQ2: do libraries that use Google Analytics implement the available privacy-protection measures?

We conducted three further tests: to measure the use of Google Analytics and Google Tag Manager across our study population; whether libraries had implemented a secure HTTPS connection between their websites and Google Analytics and/or Google Tag Manager; and whether libraries had activated Google Analytics IP anonymization. Our research results demonstrate that at a minimum 88 percent of the 279 academic library websites have implemented Google Analytics and/or Google Tag Manager, yet only 1 percent connect securely to Google via HTTPS, and only 14 percent have implemented Google Analytics IP anonymization. No library in our study activated both measures: HTTPS between their servers and Google's servers, and IP anonymization.

In the face of these results, it is clear that libraries must take additional steps to ensure that our website practices are consistent with the professional library values of privacy and intellectual freedom as articulated through organizations such as the ALA, CNI, NISO, IFLA and the Library Freedom Project. Of the library websites in our study population, many offer secure web connections for users, but most do not enforce that secure connection, and some even force users into a non-secure connection without their informed consent. And of the many websites that have implemented Google Analytics, most have done so using non-secure web connections and without activating the available privacy-protection feature of IP anonymization.

Recommendations for practice

Major library organizations such as ALA, CNI, IFLA and NISO have articulated a professional set of principles that guide our work. These principles include privacy and

intellectual freedom. However, results from our study show that the web analytics practices of many academic libraries are not in line with our profession's stated values. In order to realign toward privacy, we offer a set of practical recommendations for building more privacy-oriented library websites. Indeed, building websites that better protect user privacy can help libraries remain trusted sources of information and places of inquiry. We present our study results so that additional motivation can be generated toward building a more private web: by implementing HTTPS and automatically redirecting users to that secure URL; and by enhancing the privacy practices around Google Analytics and Google Tag Manager, which are so prevalent on library websites. We offer five recommended practices for enhancing user privacy on the web:

(1) configure library web servers to use permanent redirects (301) to HTTPS using SSL certificates provided by trusted certificate authorities;

(2) implement IP anonymization for Google Analytics;

(3) provide user education related to online privacy;

(4) obtain informed consent from users; and

(5) conduct risk/benefit analyses when using third-party service providers.

These recommendations will be further explained in the following sections. The recommendations have been developed through a synthesis of relevant literature and our research results.

Library webserver HTTPS

HTTPS is a vital privacy-protecting mechanism. By providing secure web connections for our users, we can help protect their online behavior from data leakage and provide opportunity for informed consent before surveillance occurs. Implementing HTTPS can also assist with commercial search engine discovery, as Google announced in 2014 that it would begin to favor secure websites over those that are insecure (Askey and Arlitsch, 2015). Library-led efforts such as the Library Freedom Project and its Digital Privacy Pledge are raising awareness of web privacy and the value of HTTPS[23]. The most effective method for protecting privacy calls for web servers to support ubiquitous encryption across all their domains, including all subdomains (Sivakorn *et al.*, 2016). Tools to help implement and evaluate HTTPS include Open SSL and HTTPS Everywhere[24]. Once HTTPS has been implemented, libraries can go further by ensuring that connections to their web servers occur securely via HTTPS, even when a connection is initiated through an insecure HTTP connection request. This can be accomplished by applying mechanisms that force secure connections via an HTTP 301 Redirect, which are available for WordPress[25], Apache[26], Drupal[27] and other major web publishing platforms (the precise process for implementing and enforcing HTTPS will vary according to web server platform and configurations). For the many libraries that use Google Analytics, forcing HTTPS provides the added privacy benefit of ensuring that user data transferred between the library's servers and Google's servers occurs over a secure connection.

IP anonymization for Google Analytics

Internet protocol Anonymization, also known as IP masking, is a customization to the Google Analytics tracking code that changes how Google uses and stores the IP addresses of website users[28]. This setting gives website owners using Google Analytics the option to tell Google to use only a portion of a user's IP address, thus allowing for geolocation without identification of individual users or their activity[29]. In effect, anonymizing the IP address helps protect specific identification of library website users.

User education

The privacy landscape is shifting quickly as networked technologies experience widespread development and adoption. Users want more control over tracking, though they are often unsure how to protect themselves or are distrustful of readily-available tools (Melicher *et al.*, 2015). As institutions for public good, libraries can help users understand the privacy implications of the contemporary web and, where possible, libraries can provide realistic means by which users can mitigate privacy threats. This includes informing users of privacy-based search engines such as StartPage and Duck Duck Go, IP address obfuscation through Tor relays (Acar *et al.*, 2014; Macrina, 2015a; Huang and Bashir, 2016), library workshops that advocate and educate for privacy-related topics (Gressel, 2014; Macrina, 2015b), emerging standards such as the Tracking Preference Expression (Do Not Track – DNT)[30], and independent, third-party browser tools that can help mitigate tracking, such as Disconnect and Better[31]. With third-party tracking so prevalent and sophisticated, browser add-ons and extensions are imperfect tools for ensuring privacy (Libert, 2015; Merzdovnik *et al.*, 2017; Starov and Nikiforakis, 2017), but they do add some measure of protection.

Informed consent

Informed consent increases transparency with users regarding library web tracking and privacy practices. When there is a lack of clear communication around web tracking, we can compromise the privacy of library users. Certain mechanisms currently exist to help educate users in context, such as cookie-consent notifications. This approach represents a feasible option, in that it can effectively inform users that traffic is being monitored by a third-party (Shih *et al.*, 2015). A cookie-consent notification on library websites could include, for example, a set of call-to-action buttons for accepting or declining the presence of cookies, along with follow-up links to a library's privacy policy page. Privacy policies are also useful tools for informing users about web tracking (Magi, 2007; Nichols Hess *et al.*, 2015; Kritikos and Zimmer, 2017) and for building trust with users (Aïmeur *et al.*, 2016). Privacy policies should be visible on the website and written in clear, specific language (Capistrano and Chen, 2015).

Risk/benefit analysis

Lastly, libraries and their users will benefit from a periodic risk/benefit analysis of third-party services, including analytics services such as Google Analytics. With continued projections for declining public funding for higher education, academic libraries face the pressure of demonstrating value by measuring and assessing the use of services and resources (Saunders, 2015; Alamuddin *et al.*, 2016). Google Analytics is a powerful tool for assessment, yet its connection to the vast network of third-party web trackers threatens to compromise the core library value of intellectual freedom. As Zimmer (2013, p. 56) remarks, "Libraries should minimize the use of Web cookies, bugs, and other tracking technologies." Other non-commercial web analytics services, such as Piwik, present alternatives to Google Analytics (Chandler and Wallace, 2016), although their use may be accompanied by increased administrative overhead to the library. Through a critical examination of the usage of these services, it is possible to balance the risks and the benefits to both the library and users.

Limitations and future directions

Our study design required explicit presence of the Google Analytics tracking script embedded into the homepage HTML code. This requirement produces conservative results; for instance, we discovered one case where the use of Google Tracking code was not obvious because the script was encapsulated in a separate file using a non-standard naming convention unique to

the website. This type of non-standard Google Tracking code implementation was not accounted for in our study results because our tests could not systematically identify it.

Further studies should examine the prevalence, behavior and privacy impact of additional trackers present on library websites, including those from Google Tag Manager, Google AdSense and third-party library vendors. User expectation studies will help provide additional depth and context to this research (Anton *et al.*, 2010; Lin *et al.*, 2012).

Our current study was limited to libraries found to have membership in ARL, DLF and OCLC-RLP on March 4 and 5, 2016. Future studies into HTTPS and Google Analytics can expand the study population to include academic libraries outside those membership organizations, public libraries, tribal college libraries and special libraries. Ultimately, this research is just a starting point for understanding the breadth and depth of third-party analytics tracking on library websites. We expect future research to investigate this topic to a greater extent, both in terms of the study population and the technical analysis.

Conclusion

As a profession with a long-held value of intellectual freedom, libraries should act to protect privacy on the web for their users. The ALA, CNI, NISO, IFLA and the Library Freedom Project all champion web privacy, yet the actual practices of library websites are not in alignment with their stated values of privacy. Results from our empirical study indicate that many libraries undermine user privacy by not offering secure connections to their websites, and that most libraries have implemented the third-party analytics service Google Analytics – which potentially exposes users to data leakage – but have not activated the available privacy-protection features of this tool. We conclude by offering five practical recommendations for enhancing user privacy: HTTPS, IP anonymization, user education, informed consent and risk/benefit analysis.

Data set availability

Our research data set is based on information gathered from 448 unique URLs associated with 279 North American and international academic library organizations that have membership in one or more of the following organizations: ARL, DLF or the OCLC-RLP. The following Web server responses were recorded for each URL: error response, redirect response, resolved URL, HTML cache and request protocol (HTTP or HTTPS). The data set also includes results of analysis of each URL's cached HTML used to locate and evaluate the JavaScript snippets associated with third-party web analytics trackers. The Python scripts used to make these analyses are included in the data set. The research data set and software is available through Zenodo at https://doi.org/10.5281/zenodo.1323403.

Acknowledgments

O'Brien conceived of the study, led the research design, collected data, performed the analyses, interpreted the results and critically revised the manuscript. Young participated in the research design, interpreted the results and drafted the manuscript. Arlitsch participated in the research design, interpreted the results and critically revised the manuscript. Benedict collected data, performed the analyses, interpreted the result, and critically revised the manuscript. All authors read and approved the final manuscript.

This research was conducted as part of a grant generously funded by the Institute of Museum and Library Services (IMLS). "Measuring Up: Assessing Accuracy of Reported Use and Impact of Digital Repositories" was a three-year research project led by Montana State University, in partnership with the Association of Research Libraries, OCLC Research, and the University of New Mexico (IMLS Log Number: LG-06-14-0090-14). The grant proposal narrative may be found at http://scholarworks.montana.edu/xmlui/handle/1/8924.

Notes

1. https://support.google.com/analytics/answer/2763052?hl=en; https://support.google.com/analytics/answer/2905384?hl=en (accessed December 9, 2016).
2. https://support.google.com/analytics/answer/181881?hl=en (accessed December 9, 2016).
3. https://webmasters.googleblog.com/2014/08/https-as-ranking-signal.html; https://support.google.com/webmasters/answer/6073543
4. www.ala.org/advocacy/privacyconfidentiality
5. www.ala.org/advocacy/intfreedom/librarybill/interpretations/privacy
6. www.ala.org/advocacy/privacyconfidentiality/toolkitsprivacy/privacy-and-confidentiality-library-core-values
7. www.cni.org/news/privacy-in-the-age-of-analytics-executive-roundtable-report-available
8. www.niso.org/apps/group_public/download.php/15863/NISO%20Consensus%20Principles%20on%20Users%C2%92%20Digital%20Privacy.pdf
9. www.ifla.org/node/9803
10. https://libraryfreedomproject.org/ourwork/digitalprivacypledge/
11. www.arl.org/membership/list-of-arl-members
12. www.oclc.org/research/partnership/roster.html
13. www.diglib.org/members/
14. "Benefits of Membership | Association of Research Libraries®|ARL®." Accessed November 17, 2016, available at: www.arl.org/membership/benefits#.WC3pJ9ym3i8; "The OCLC Research Library Partnership." Accessed November 17, 2016, available at: http://www.oclc.org/research/partnership.html; "About the Digital Library Federation." Accessed November 17, 2016, available at: https://www.diglib.org/about/
15. https://doi.org/10.5281/zenodo.1323403
16. https://github.com/imls-measuring-up/library-privacy
17. Google. "Check If a Web Page Uses Google Analytics." Google Analytics Help, 2016. https://support.google.com/analytics/answer/1032399
18. Google. "Google Tag Manager: Quick Start Guide." 2017. https://developers.google.com/tag-manager/quickstart
19. Google. "Analytics.js Field Reference|Analytics for Web (Analytics.js)|Force SSL." Google Developers, accessed September 23, 2016. Available at: https://developers.google.com/analytics/devguides/collection/analyticsjs/field-reference
20. At the time of this writing, we could not locate any Google Tag Manager documentation describing features related to protecting patron privacy. Therefore, the presence of Google Tag Manager on a website indicates that no privacy-protection measures are in place
21. Google. "Analytics.js Field Reference|Analytics for Web (Analytics.js)|Force SSL." Google Developers, accessed September 23, 2016. Available at: https://developers.google.com/analytics/devguides/collection/analyticsjs/field-reference
22. https://en.wikipedia.org/wiki/Man-in-the-middle_attack
23. https://libraryfreedomproject.org/ourwork/digitalprivacypledge/
24. www.openssl.org; https://www.eff.org/https-everywhere
25. https://wordpress.org/plugins/wp-force-ssl/; https://wordpress.org/plugins/really-simple-ssl/
26. https://wiki.apache.org/httpd/RedirectSSL

27. www.drupal.org/https-information
28. https://support.google.com/analytics/answer/2905384?hl=en; https://support.google.com/analytics/answer/2763052
29. https://support.google.com/analytics/answer/6004245
30. www.w3.org/2011/tracking-protection/drafts/tracking-dnt.html
31. https://disconnect.me; https://better.fyi

References

Acar, G., Eubank, C., Englehardt, S., Juarez, M., Narayanan, A. and Diaz, C. (2014), "The web never forgets: persistent tracking mechanisms in the wild", *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, *New York, NY*, pp. 674-689.

Aïmeur, E., Lawani, O. and Dalkir, K. (2016), "When changing the look of privacy policies affects user trust: an experimental study", *Computers in Human Behavior*, Vol. 58 No. Supplement C, pp. 368-379.

Alamuddin, R., Kurzweil, M. and Rossman, D. (2016), "Higher ed insights: results of the spring 2016 survey", Ithaka S+R, available at: https://doi.org/10.18665/sr.284439 (accessed December 6, 2016).

Almind, T.C. and Ingwersen, P. (1997), "Informetric analyses on the world wide web: methodological approaches to 'webometrics'", *Journal of Documentation*, Vol. 53 No. 4, pp. 404-426.

Anton, A.I., Earp, J.B. and Young, J.D. (2010), "How internet users' privacy concerns have evolved since 2002", *IEEE Security Privacy*, Vol. 8 No. 1, pp. 21-27.

Arlitsch, K. and OBrien, P.S. (2013), *Improving the Visibility and Use of Digital Repositories Through SEO: A LITA Guide*, ALA Neal-Schuman, Chicago, IL.

Askey, D. and Arlitsch, K. (2015), "Heeding the signals: applying web best practices when Google recommends", *Journal of Library Administration*, Vol. 55 No. 1, pp. 49-59.

Barba, I., Cassidy, R., Leon, E.D. and Williams, B.J. (2013), "Web analytics reveal user behavior: TTU libraries' experience with Google analytics", *Journal of Web Librarianship*, Vol. 7 No. 4, pp. 389-400.

Bashir, M.A., Arshad, S., Robertson, W. and Wilson, C. (2016), "Tracing information flows between ad exchanges using retargeted ads", *25th USENIX Security Symposium*, *Austin, TX*, pp. 481-496.

Björneborn, L. and Ingwersen, P. (2001), "Perspectives of webometrics", *Scientometrics*, Vol. 50 No. 1, pp. 65-82.

Björneborn, L. and Ingwersen, P. (2004), "Toward a basic framework for webometrics", *Journal of the American Society for Information Science and Technology*, Vol. 55 No. 14, pp. 1216-1227.

Brotsky, S.R. and Giles, D. (2007), "Inside the 'pro-ana' community: a covert online participant observation", *Eating Disorders*, Vol. 15 No. 2, pp. 93-109.

Capistrano, E.P.S. and Chen, J.V. (2015), "Information privacy policies: the effects of policy characteristics and online experience", *Computer Standards & Interfaces*, Vol. 42 No. Supplement C, pp. 24-31.

Caro, A. and Markman, C. (2016), "Measuring library vendor cyber security: seven easy questions every librarian can ask", *The Code4Lib Journal*, No. 32, available at: http://journal.code4lib.org/articles/11413 (accessed September 18, 2016).

Chandler, A. and Wallace, M. (2016), "Using Piwik instead of Google analytics at the Cornell university library", *The Serials Librarian*, Vol. 71 Nos 3–4, pp. 173-179.

Cohen, R.A. and Thorpe, A. (2015), "Discovering user behavior: applying usage statistics to shape frontline services", *The Serials Librarian*, Vol. 69 No. 1, pp. 29-46.

Conrad, S. (2015), "Using Google tag manager and google analytics to track DSpace metadata fields as custom dimensions", *Code4Lib Journal*, No. 27, available at: http://journal.code4lib.org/articles/10311 (accessed October 5, 2016).

Englehardt, S. and Narayanan, A. (2016), "Online tracking: a 1-million-site measurement and analysis", *Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security*, *Vienna*, *October*.

Fagan, J.C. (2014), "The suitability of web analytics key performance indicators in the academic library environment", *The Journal of Academic Librarianship*, Vol. 40 No. 1, pp. 25-34.

Falahrastegar, M., Haddadi, H., Uhlig, S. and Mortier, R. (2016), "Tracking personal identifiers across the web", *Passive and Active Measurement, presented at the International Conference on Passive and Active Network Measurement, Springer*, *Cham*, pp. 30-41.

Farney, T. (2016), "Google analytics and Google tag manager", *Library Technology Reports*, Vol. 52 No. 7, pp. 1-42.

Gardner, G. and Groover, M. (2015), "Web privacy in practice: assessing internet security and patron privacy in North American public libraries", November 12, available at: http://macsphere.mcmaster.ca/handle/11375/19016 (accessed October 21, 2016).

Garoogian, R. (1991), "Librarian/patron confidentiality: an ethical challenge", *Library Trends*, Vol. 40 No. 2, pp. 216-233.

Gressel, M. (2014), "Are libraries doing enough to safeguard their patrons' digital privacy?", *The Serials Librarian*, Vol. 67 No. 2, pp. 137-142.

Hess, K. (2012), "Discovering digital library user behavior with Google analytics", *The Code4Lib Journal*, No. 17.

Huang, H.-Y. and Bashir, M. (2016), "The onion router: understanding a privacy enhancing technology community", *Proceedings of the Association for Information Science and Technology*, Vol. 53 No. 1, pp. 1-10.

Johnston, S.D. (2000), "Rethinking privacy in the public library", *International Information & Library Review*, Vol. 32 Nos 3–4, pp. 509-517.

Kalavri, V., Blackburn, J., Varvello, M. and Papagiannaki, K. (2016), "Like a pack of wolves: community structure of web trackers", in Karagiannis, T. and Dimitropoulos, X. (Eds), *Passive and Active Measurement*, Springer International Publishing, Cham, pp. 42-54.

Krishnamurthy, B. and Wills, C. (2009), "Privacy diffusion on the web: a longitudinal perspective", *Proceedings of the 18th International Conference on World Wide Web, ACM*, *New York, NY*, pp. 541-550.

Kritikos, K.C. and Zimmer, M. (2017), "Privacy policies and practices with cloud-based services in public libraries: an exploratory case of bibliocommons", *Journal of Intellectual Freedom and Privacy*, Vol. 2 No. 1, pp. 23-37.

Kurian, G.T. (2013), *The AMA Dictionary of Business and Management*, AMACOM, New York, NY.

Lerner, A., Simpson, A.K., Kohno, T. and Roesner, F. (2016), "Internet jones and the raiders of the lost trackers: an archaeological study of web tracking from 1996 to 2016", *Presented at the 25th USENIX Security Symposium (USENIX Security 16)*, *Austin, TX, August 10-12*, available at: www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/lerner (accessed September 17, 2016).

Libert, T. (2015), "Exposing the invisible web: an analysis of third-party HTTP requests on 1 million websites", *International Journal of Communication*, Vol. 9, available at: http://ijoc.org/index.php/ijoc/article/view/3646 (accessed October 5, 2016).

Lin, J., Amini, S., Hong, J.I., Sadeh, N., Lindqvist, J. and Zhang, J. (2012), "Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing", *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, *New York, NY*, pp. 501-510.

Macrina, A. (2015a), "Accidental technologist: the tor browser and intellectual freedom in the digital age", *Reference & User Services Quarterly*, Vol. 54 No. 4, pp. 17-20.

Macrina, A. (2015b), *Privacy Advocacy in Libraries in the Age of Mass Surveillance*, LACUNY Institute, New York, NY, available at: http://academicworks.cuny.edu/lacuny_conf_2015/7 (accessed September 25, 2017).

Magi, T.J. (2007), "The gap between theory and practice: a study of the prevalence and strength of patron confidentiality policies in public and academic libraries", *Library & Information Science Research*, Vol. 29 No. 4, pp. 455-470.

Melicher, W., Sharif, M., Tan, J., Bauer, L., Christodorescu, M. and Leon, P.G. (2015), "(Do not) track me sometimes: users' contextual preferences for web tracking", *Proceedings on Privacy Enhancing Technologies*, Vol. 2016 No. 2, pp. 135-154.

Merzdovnik, G., Huber, M., Buhov, D., Nikiforakis, N., Neuner, S., Schmiedecker, M. and Weippl, E. (2017), "Block me if you can: a large-scale study of tracker-blocking tools", *2017 IEEE European Symposium on Security and Privacy (EuroS P)*, pp. 319-333.

Million, A.C. and Fisher, K.N. (1986), "Library records: a review of confidentiality laws and policies", *Journal of Academic Librarianship*, Vol. 11 No. 6, pp. 346-349.

Narayanan, A. and Reisman, D. (2017), "The Princeton web transparency and accountability project", in Cerquitelli, T., Quercia, D. and Pasquale, F. (Eds), *Transparent Data Mining for Big and Small Data*, Springer, Cham, pp. 45-67.

Naylor, D., Finamore, A., Leontiadis, I., Grunenberger, Y., Mellia, M., Munafò, M., Papagiannaki, K. *et al.* (2014), "The cost of the 'S' in HTTPS", *Proceedings of the 10th ACM International on Conference on Emerging Networking Experiments and Technologies*, *New York, NY*, pp. 133-140.

Nichols Hess, A., LaPorte-Fiori, R. and Engwall, K. (2015), "Preserving patron privacy in the 21st century academic library", *The Journal of Academic Librarianship*, Vol. 41 No. 1, pp. 105-114.

Olejnik, L., Castelluccia, C. and Janc, A. (2012), "Why Johnny can't browse in peace: on the uniqueness of web browsing history patterns", presented at the 5th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs 2012), available at: https://hal.inria.fr/hal-00747841/document (accessed October 5, 2016).

Punch, K. (2014), *Introduction to Social Research: Quantitative & Qualitative Approaches*, 3rd ed., SAGE, Los Angeles, CA.

Roesner, F., Kohno, T. and Wetherall, D. (2012), "Detecting and defending against third-party tracking on the web", *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation, USENIX Association*, *Berkeley, CA*, pp. 12-12.

Sar, R.K. and Al-Saggaf, Y. (2013), "Propagation of unintentionally shared information and online tracking", *First Monday*, Vol. 18 No. 6, available at: http://firstmonday.org/ojs/index.php/fm/article/view/4349 (accessed September 18, 2016).

Saunders, L. (2015), "Academic libraries' strategic plans: top trends and under-recognized areas", *The Journal of Academic Librarianship*, Vol. 41 No. 3, pp. 285-291.

Schelter, S. and Kunegis, J. (2016), "Tracking the trackers: a large-scale analysis of embedded web trackers", *Tenth International AAAI Conference on Web and Social Media*, *Cologne*, *May 17-20*, available at: www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13024 (accessed September 20, 2016).

Shih, F., Liccardi, I. and Weitzner, D. (2015), "Privacy tipping points in smartphones privacy preferences", *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, *New York, NY*, pp. 807-816.

Sivakorn, S., Keromytis, A.D. and Polakis, J. (2016), "That's the way the cookie crumbles: evaluating HTTPS enforcing mechanisms", *Proceedings of the 2016 ACM on Workshop on Privacy in the Electronic Society*, *New York, NY*, pp. 71-81.

Stanley, R. and McLaren, S. (2007), "Ethical issues in health and social care research", in Leathard, A., Goodinson-McLaren, S. and McClaren, S. (Eds), *Ethics: Contemporary Challenges in Health and Social Care*, Policy Press, Bristol, p. 314.

Starov, O. and Nikiforakis, N. (2017), "Extended tracking powers: measuring the privacy diffusion enabled by browser extensions", *Proceedings of the 26th International Conference on World Wide Web*, *Republic and Canton of Geneva*, pp. 1481-1490.

Taylor, S.J., Bogdan, R. and DeVault, M.L. (2016), *Introduction to Qualitative Research Methods: A Guidebook and Resource*, 4th ed., John Wiley & Sons, Hoboken, NJ.

Togia, A. and Malliari, A. (2017), "Research methods in library and information science", in Oflazoglu, S. (Ed.), *Qualitative versus Quantitative Research*, InTech, London, pp. 43-64, available at: https://doi.org/10.5772/intechopen.68749 (accessed May 29, 2018).

Yang, L. and Perrin, J.M. (2014), "Tutorials on google analytics: how to craft a web analytics report for a library web site", *Journal of Web Librarianship*, Vol. 8 No. 4, pp. 404-417.

Zimmer, M. (2013), "Patron privacy in the '2.0' era: avoiding the Faustian Bargain of library 2.0", *Journal of Information Ethics*, Vol. 22 No. 1, pp. 44-59.

Zimmer, M. (2014), "Librarians' attitudes regarding information and internet privacy", *The Library Quarterly*, Vol. 84 No. 2, pp. 123-151.

Corresponding author
Scott W.H. Young can be contacted at: swyoung@montana.edu