Historical Website Ecology

Analyzing Past States of the Web Using Archived Source Code

Helmond, A.

Publication date
2017
Document Version
Author accepted manuscript
Published in
Web 25

Helmond, Anne. forthcoming. “Historical Website Ecology. Analyzing Past States of the Web Using Archived Source Code.” In *Web 25: Histories from the First 25 Years of the World Wide Web*, edited by Niels Brügger. New York: Peter Lang Publishing.

Historical website ecology
Analyzing past states of the web using archived source code

Anne Helmond, University of Amsterdam

In this chapter I offer a historical perspective on the changing composition of a website over time. I propose to see the website as an ecosystem through which we can analyze the larger techno-commercial configurations that websites are embedded in. In doing so, I reconceptualize the study of websites as historical website ecology. The website’s ecosystem can be detected by examining the source code in which a website’s connections with third parties have become inscribed. If archived, this provides a way to examine changes in a website’s ecosystem as a way to transformations in the techno-commercial configurations of the web through the changing composition of a website. Moving the site of analysis from the content of websites to the context of websites opens up new areas for web historical research. Focusing on the archived source code of websites does not only enable analyzing web technologies used to construct them, which can tell us something about the web’s underlying infrastructure, providing insights to how the web is built and how websites are connected, but can also serve as a means to investigate the web’s economic underpinnings, to understand the business models of websites and third parties and trace the economically valuable data flowing between them. In this chapter I take a contextual approach to historical website analysis by viewing the website as an environment that is inhabited and shaped by third parties such as social media platforms, advertisers, analytics companies and content-delivery networks, embedding the website in various technological and commercial relations with these actors. This shift from website

content to website context is what I refer to as the website's ecosystem as a way to study changes in the techno-commercial environment of the web.

I position the historical study of a website's ecosystem as a contribution to web historiography, which is concerned with writing histories of the web (Ankerson, 2012; Brügger, 2009, 2013; Foot & Schneider, 2010). To operationalize my contribution, I turn to web archives as important tools for web historians to uncover previous states of the web. I argue that while the web archive of the Internet Archive Wayback Machine focuses on snapshots of single websites (Brügger 2009; Rogers 2013; Ben-David and Huurdeman 2014) the source code of these snapshots contains valuable information about a website's relations with third parties that we can employ for the reconstruction of historical website ecosystems. That is, I propose treating the source code as a demarcation object that determines the dynamic interrelations between websites and external actors, by focusing on the code snippets of third-party objects that enable these connections. Whilst archived source code may be used to analyze different aspects over time, such as a website's relations with social media platforms through sharing features and login systems, the web technologies used, as well as insights into the business practices of websites through their use of online advertising and analytics, in this chapter I focus on one particular technological type of third-party connection that is established through tracking mechanisms embedded in websites. Trackers are of particular relevance because of the increasing interest and use of (big) data collected on people visiting websites and the changes in techniques from third parties to collect data on external websites (Turow, 2011). In a case study on trackers on the New York Times (NYT) website I examine how we can employ archived source code to reconstruct the historical tracking ecologies the NYT website has been embedded in between 1996-2011.

Website ecology

Mayer and Mitchell claim that websites are increasingly being shaped by third-party content and functionality (2012). For example, webmasters can use social media platforms to embed sharing functionality with social buttons, and advertising servers to display dynamically-generated personalized ads to generate income. In these scenarios,

the website is no longer a self-contained unit but has become informed and molded by these other actors on the web.

Thus, I would like to introduce the notion of *website ecology*, that is the study of the complex socio-technical relations between websites, users, social media platforms, tracking companies, and other actors in the website's environment. In doing so, I draw parallels with media ecology, defined by Neil Postman as "the study of media as environments […] their structure, content, and impact on people" (1970). Shifting media ecology's focus away from studying the effects of media on people towards the materiality of these media environments, Matthew Fuller argues for understanding media ecology as "the massive and dynamic interrelation of processes and objects, beings and things, patterns and matter" (2005). Here, I am particularly interested in the "softwarization" of media (Berry, 2012; Manovich, 2013, p. 5), a term used to reconceptualize media ecology after our "media becomes software" (Manovich, 2013, p. 156). David Berry employs the term ecology as "as a broad concept related to the environmental habitus of both human and non-human actors" (2012) to describe our current media system as a "computational ecology" which is comprised of distinct "software ecologies" (2012).

Following these authors, I draw from ecology in a similar manner to analyze changes in the composition of the web by studying the relations between a website and its environment. Website ecology looks at the dynamic and shifting relations between websites and third parties, which do not only become interconnected through users' web activities such as linking pages, but also through software features such as social buttons and data connections created by trackers on websites. An ecological approach to understanding websites allows for the analysis of websites as dynamic spaces where these complex relations between users, websites and third parties such as tracking companies and advertisers get encoded.

Studying websites and their environments through the source code

Previous approaches to studying the website in its networked environment have focused on how websites establish relations with other websites through linking, therewith embedding the website in a hyperlink network (Park, 2003; Rogers, 2002). Greg Elmer and Ganaele Langlois describe these approaches as "Web 1.0 methods focused on mapping

hyperlink networks" (2013, p. 43) to analyze the connections between websites and the networks they form. Previously, Elmer has called for detecting new indicators of networking and to develop

> a broader vision for the analysis of web code, expanding beyond the mapping of HREF tags (hyperlink code) toward an understanding of the larger structure and deployment of all web code and content (including text, images, met tags, robot.txt commands and so on) (2006, p. 9).

In this chapter I contribute a new approach that examines the source code for scripts that create connections with third parties. Web 2.0, or now commonly referred to as the social web, is characterized by new forms of networked connectivity, Elmer and Langlois argue, which move beyond the hyperlink and which require new methods to map these new types of connections (2013, p. 44). They outline the "building blocks" of what they refer to as Web 2.0 "cross platform based methods" in which they trace objects across platforms to detect their channels of circulation and analyze the different types of relationships that they form (2013, p. 45). Here, I build on Elmer and Langlois' idea of cross platform analysis with a novel method that traces cross connections from within a website by employing the embedded third-party objects.

A cross platform approach shifts the attention away from the hyperlink as the prime connection mechanism towards other web objects that create interactions between a user, a website and its ecosystem. Of particular interest here are the objects that are not immediately visible in the front-end, the end-user interface of the website, and which create relations with other actors on the web such as trackers. Roesner et al. define a (third-party) tracker as "a website (like doubleclick.net) that has its tracking code included or embedded in another site (like cnn.com)" to "identify and collect information about users" (2012, p. 12).

Web bugs, beacons and other types of trackers embed the website in larger techno-commercial configurations on the web by establishing relations between websites and advertising networks, analytics companies and market research companies, amongst others. Detecting these relations requires moving beyond the user interface in order to detect the traces of these dynamic relations by engaging with the materiality of a website, the source code.

In this chapter, I draw from European media ecology which emphasizes the materiality of code and software of our contemporary media environment (Berry, 2012; Fuller, 2005; Goddard & Parikka, 2011). In this view, the source code of a website forms the object of the study of website ecology. The website's source code provides the material in which the relations with other actors become inscribed through dynamic third-party content, objects and features. This follows a perspective advocated by a number of authors (Ankerson, 2012; Brügger, 2008; De Souza, Froehlich, & Dourish, 2005; Marino, 2014) who engage with the materiality of new media by pointing to the source code as an important entry point for web historical analysis.

De Souza et al. propose to see the source code as "a social and technical artifact" in which aspects of software development have become inscribed (2005, p. 197). They draw from Latour's notion of inscription (1999)[1] to refer to "a process through which social practice and technological artifacts become inextricably intertwined" (2005, p. 197). They see "software artifacts as pure inscriptions" that can be used "to uncover the structure of software projects" and their development processes (2005, p. 197), an approach they refer to as "an 'archeology' of software processes" (2005, p. 206). Similarly, Megan Ankerson turns to the traces of software to engage with "the culture of software in constructing histories of the web" thereby bringing a "software studies lens to web historiography" (2009, p. 195). I build on De Souza et al.'s and Ankerson's approaches by seeing the archived source code as a document in which relations with third parties become inscribed and which that can be used to reconstruct historical techno-commercial configurations on the web.

Techno-commercial environments: Tracking ecosystems

Trackers in the form of beacons and analytics can be implemented by webmasters to monitor the functioning of their websites or to collect data about their visitors. However, trackers can also come as a by-product of third-party functionality such as the Facebook Like Button on external websites which tracks Facebook users and non-Facebook-users

[1] Latour defines inscription as "a general term that refers to all the types of transformations through which an entity becomes materialized into a sign, an archive, a document, a piece of paper, a trace" (1999, p. 306).

across the web (Gerlitz & Helmond, 2013). Webmasters that employ website analytics, advertisements or social buttons therewith—intentionally or unintentionally—embed the website and its visitors in a tracking ecosystem.

Since the early days of the web, banner ads, cookies and other tracking web objects have been an integral part of the web. In October 1994, HotWired placed the first banner ads on its website which marked an important turning point for the web (D'Angelo, 2009). Another important development was the creation of ad networks connecting advertisers to webmasters (Gehl, 2014, p. 104). With the rise of ad servers to track and monitor ads on third-party websites, trackers have become an integral part of the interactions between websites and their ecosystem (Mayer & Mitchell, 2012). In addition, the number of advertising and analytical companies and type of tracking mechanisms has increased over the years as the result of an growing interest in the collection of user data for advertising purposes (Turow, 2011). Next, I will discuss what this means for the changing composition of the website.

The website as an assembled unit

In the early days of the web, often referred to as Web 1.0, websites were considered to be fairly self-contained units since most content was stored on the same server (Mayer & Mitchell, 2012; Song, 2010, p. 251). Within Web 2.0, now more commonly referred to as the social web, websites are considered to be increasingly entangled in a networked context and shaped by third-party content and dynamically-generated functionality (Gehl, 2014, p. 103; Liu, 2004; Mayer & Mitchell, 2012).

This increasing modularity of websites is key to the changing nature of the website which, Alan Liu contends, can be understood as shifting from web pages as independent units to webpages which are filled with content from external databases (2004). Robert Gehl similarly argues that in Web 2.0 a website is assembled from third-party sources:

> a website is a 'mash-up' of top-down, incrementally altered architecture, bottom-up user participation and processing, and the lateral insertion of advertising, creating a coherent visual artifact out of these different streams (2014, p. 103).

The website can be seen as an assemblage of modular elements that on the one hand enable interactions with other actors on the web and on the other hand permeate or redraw the

boundaries of the website by setting up data channels for the exchange of content and data stored in external databases. Next, I will show that the website as an assembled object provides an important entry point for analyzing historical website ecosystems through web archives.

The detachment of the website ecosystem

To study previous states of the web, the web historian needs access to historical material which can be found in web archives such as the Internet Archive Wayback Machine (IAWM). However, web historian Niels Brügger argues, the archiving process actively shapes and determines how a website is archived and therefore what kind of reconstruction or historical analysis is possible (2009, p. 126). He argues that the effect of the arching process is that "in practice the website is almost always the basic unit in a web archive" (2013, p.757). Brügger defines the website as

> a coherent textual unit that unfolds in one or more interrelated browser windows, the coherence of which is based on semantic, formal and physically performative interrelations (2009, p. 126).

Web historians Foot and Schneider delineate websites as "groups of pages sharing a common portion of their URL" (2010, p. 69) where web pages are seen as "groups of elements assembled by a producer and displayed upon request to a server" (2010, p. 69). These definitions of the website as an object of study focus on the visible rendering of the website as a coherent yet assembled object. What we see in the archiving process is that archived website becomes detached from its larger context.

This may be seen in the archived websites in the IAWM, one of the largest available web archives. The IAWM, Richard Rogers argues, lends itself to "single-site histories" or "website biographies" as one accesses the archive by entering a single URL, the website's domain name (2013, p. 66). The focus on the single website shows how in the process of archiving, the website has been separated from its ecosystem. Web archives often privilege the content of websites over the search engine results they are part of, their Alexa rankings or their statistics, amongst others (Rogers, 2013, p. 63). Thus, in the archiving process the website is detached from the techno-social context it resides in (Weltevrede, 2009, p. 84).

While the website is often the main unit within web archives (Brügger, 2013, p. 756), the archived website's source code also contains elements that can be employed to "uncover parts of the web that were not preserved" (Samar, Huurdeman, Ben-David, Kamps, & de Vries, 2014, p. 1199). In what follows next I develop a novel method that moves beyond the single-site history by employing the code snippets of an archived website to reconstruct a website's ecosystem. My method addresses the conceptual and practical problem of the website as a bounded object which has troubled web archiving theorists (cf. Brügger, 2009; cf. Schneider & Foot, 2004). In shifting the focus from the content of the archived website to the code, different analytical opportunities present themselves.

Analyzing historical website ecosystems

In the source code of archived websites we can find the traces, the code snippets, of web objects such as trackers and other third-party content. Following Matthew Kirschenbaum (2003) and Megan Sapnar Ankerson (2009), I use the word traces to refer to the material evidences of software, in this case the tracker code. While trackers or tracking objects may not be archived their code traces allow for reconstructing the tracking networks that websites have been embedded in. Previous approaches in using aspects of a website's source code to study its environment include outlinks to other websites as one way to move beyond the single-site history. The HTML code for hyperlinks in an archived website's source code enables the reconstruction of past hyperlink networks through a historical hyperlink analysis.[2] Even though these websites may not have been archived themselves, the outlinks pointing to them allows for "conjuring up" these websites (Stevenson, 2010) and map past states of the web or the blogosphere using the IAWM (Ammann, 2011; Ben-David, 2012; Stevenson, 2010; Weltevrede & Helmond, 2012). Following Elmer (2006) and Elmer and Langlois' (2013) call for employing other web objects for networking, I

[2] Despite the fact that web archives are often incomplete, the proposed method does not require the linked website to be archived as well. The mere presence of the hyperlink pointing to it can be used to map the historical hyperlink network of a blogosphere, see the work of (Ammann, 2011; Ben-David, 2012; Stevenson, 2010; Weltevrede & Helmond, 2012).

move beyond the hyperlink and focus on trackers as objects that entangle the website in a techno-commercial web environment.

Previous historical tracker studies include a longitudal study on trackers on 1200 websites between October 2005 and September 2008 (Krishnamurthy & Wills, 2009) and cookies and their (default) settings in different Netscape browser versions over time (Elmer, 2002). While web archiving has taken the first steps to attend to websites as part of hyperlink networks, little attention has been paid to the historical study of tracker networks so far.

My proposed methodology to analyze historical tracking networks builds on previous research to map tracking networks (Gerlitz & Helmond, 2013; van der Velden, 2014). In this chapter I further extend these methods to detect trackers in *archived websites* to analyze the tracking networks that websites have been embedded in over time. At the same time this methodology serves as a blueprint for developing further methods that focus on detecting and mapping other third-party objects such as social buttons in archived websites to study previous states of the web through website features and technologies to examine changes in the web's techno-commercial environments.

The methodology to create tracker networks is inspired by a digital methods approach of "repurposing" the existing analytical capacities of tools and devices for research with the web (Rogers, 2013, p. 1; Weltevrede, 2016). Many tools on the web have a methodological approach built into them to achieve a particular functionality. An example of such a tool is the browser add-on Ghostery, which

> scans the page for trackers—scripts, pixels, and other third party elements—and notifies you of the companies whose code is present on the webpages you are visiting. Usually, these trackers aren't visible, and they are often hard to find in the page source code (Ghostery, 2013).

Ghostery has an inbuilt method to detect trackers in websites that can be employed for research purposes. Instead of creating a new method or tool to find trackers, we can also repurpose the existing analytical capacities of Ghostery. Colleagues and I repurposed Ghostery and created the Tracker Tracker tool (Borra et al., 2012) to analyze and map the presence of trackers in a collection of websites.

Ghostery looks for patterns of trackers and matches them to a database of over 2000 known trackers.[3] It uses simple string matching (matching a number of characters in a code string) and regex (regular expressions) as a method to detect and match the found tracker code against their database of trackers. For example, Ghostery looks for the presence of advertiser DoubleClick on a website by examining the website's source code for known DoubleClick patterns in Ghostery's tracker database, for example [ad.doubleclick.net] or [doubleclick.net/pagead]).

The main contribution of repurposing Ghostery—by building the new Tracker Tracker tool on top of it—is to detect and map *tracker networks*. While Ghostery has been developed as a plugin to detect and block trackers on *individual websites*, the Tracker Tracker tool is able to detect trackers in *collections of websites* and to create *a network view* of websites and their trackers. The Tracker Tracker tool scans URLs and outputs the name of the website, the tracker found, the tracker pattern and the tracker type in a .csv spreadsheet and .gefx file. This latter file, a Gephi graph, also contains the relations between the trackers and the websites in the collection, based on tracker presence, and can be used to visualize the network of websites and their trackers using the graph analysis and visualization tool Gephi (Bastian, Heymann, Jacomy, & others, 2009).

Scanning large collections of websites for trackers enables the mapping of tracking networks that websites are embedded in. Such an approach no longer focuses on the relations between websites, by reconstructing a network based on the visible outlinks found on the website, but instead focuses on the invisible connections established with trackers such as central ad servers and platforms in the back-end. This approach, which looks at different devices to organize relations on the web, such as trackers, shows a specific view of the website's ecosystem. It allows for the reconstruction of a different network of connectivity operating in the back-end of a website by employing the data flows between websites and advertising, tracking and analytics companies as will be demonstrated in the following case study.

[3] See: https://www.ghostery.com/our-solutions/ghostery-browser-extention/ 1 [Accessed 10 January 2016].

Tracing the trackers using the Internet Archive Wayback Machine

In this case study I develop a novel method to map the tracking ecology of the New York Times (NYT) website over time by using the Tracker Tracker tool in combination with the Internet Archive Wayback Machine (IAWM). I use the IAWM because it provides a valuable source for web historians because of its accessibility and scope. The IAWM contains over 464 Billion URLs[4] and provides snapshots from a wide range of archived websites from 1996 until very recent.

While trackers, or the websites issuing the trackers, may no longer exist or be in use, their code snippets—the code that enables the tracking—can still be found in archived websites within the IAWM. Trackers are visible in the archived website's source code because they are either "hardcoded" into the source code or have become imprinted in the website's source code during the archival process.

An important finding is that Ghostery still detects trackers in archived websites when surfing through the IAWM with Ghostery enabled. When verifying the detected trackers, by manually comparing the source code of the archived website with the Ghostery database there is indeed a match with the found pattern [ad.doubleclick.net]. Many patterns are established on domain name [ad.doubleclick.net] or subpage level [doubleclick.net/pagead]) which means that even if the tracking technique changes, the tracker will still be detected if it is issued from the same domain or subpage. According to Ghostery's developers, trackers' issuing domains have hardly changed since they started developing the plugin, making it a suitable tool for historical research.[5] The question remains to what extent the Ghostery database contains old trackers. While the example of DoubleClick, which has existed since 1996, shows that such companies can be traced and detected in retrospect, the detection of old trackers relies on the assumption that tracker (sub)domains do not change over time. This means that the approach put forward here can only detect trackers that Ghostery currently has in its database and further research should therefore investigate old tracker patterns. Concerning new trackers, Ghostery operates a

[4] See: https://archive.org/web/ [Accessed 23 January 2016].

[5] On May 30, 2013 I visited Ghostery's office (Evidon Inc.) in New York City and interviewed Ghostery's developers about their tracker database and the plugin's technical functionality.

cumulative database and constantly adds new patterns of existing trackers to its database as well as new trackers.

What follows from the observation that Ghostery is still able to detect trackers in archived websites is that some of the functionality of the trackers resumes to exist within archived websites and that tracking companies continue to track users within the archived web. It also means that the Tracker Tracker tool—which is based on Ghostery—works with the IAWM which makes it possible to detect and analyze a website's tracking ecology over time. Next, I will detail a method to do so.

Method

The object of study is the archived NYT website—or more specifically the archived NYT front pages. The NYT was chosen due to the site's centrality as a news source, its large number of visitors per day[6] and the presence of a fair amount of trackers. The selected time frame is 1996-2011, from the first (1996) to the final (2011) full year the NYT was archived in the IAWM at the time of the case study in December 2012. In a first step I set out to collect all the Internet Archive URLs for the archived snapshots from the NYT in the IAWM.

Instead of collecting all IAWM URLs manually, I have used the Internet Archive Wayback Machine Link Ripper tool (Borra et al., 2009) to automate the process. This tool retrieves the links of a website's archived snapshots at wayback.archive.org. The input is a URL [http://www.nytimes.com/] and the output is a text file which lists the IAWM URLs—the links of the archived snapshots. In case the IAWM has archived multiple snapshots of the website per day, only the first archived version of that day is retained and listed in the resulting text file. Since the advertising and analytical techniques that are implemented by the NYT's webmasters do not change on a daily basis, this does not affect the research setup. Table 1 shows the number of URLs collected by the IAWM Link Ripper tool per year.

[6] See http://www.alexa.com/siteinfo/nytimes.com and http://www.alexa.com/topsites/category/Top/News [accessed 12 January 2013].

‘96	‘97	‘98	‘99	‘00	‘01	‘02	‘03	‘04	‘05	‘06	‘07	‘08	‘09	‘10	‘11
7	4	1	13	10	174	157	71	21	230	208	290	261	243	281	352

Table 1: The number of IAWM URLs collected for the NYT website between 1996 and 2011 per year. The URLs were retrieved by the IAWM Link Ripper tool.

In a second step I used the Tracker Tracker tool to scan the IAWM URLs—the archived snapshots—for tracking technologies. The input is the list of IAWM URLs that was compiled in the previous step and the output shows the detected trackers per URL. The result file can be downloaded from the tool in CSV (spreadsheet), GEFX (Gephi) or HTML-format and contains the IAWM URLs, the tracker name, type and pattern that was detected. The type of trackers follows the categorization provided by Ghostery: Ad, Analytics, Beacon/Tracker, and Widget.[7] I then collected this detailed information in a spreadsheet. Figure 1 shows the number and type of trackers that have been detected in the archived snapshots of the NYT front page per year.[8]

[7] See: http://www.knowyourelements.com/#tab=intro [accessed 12 January 2013].
[8] There are a number of gaps in the available IAWM data from 1996-2000 and in 2004 data is missing for the months April-September which may explain the sudden “decline” in trackers in 2004.

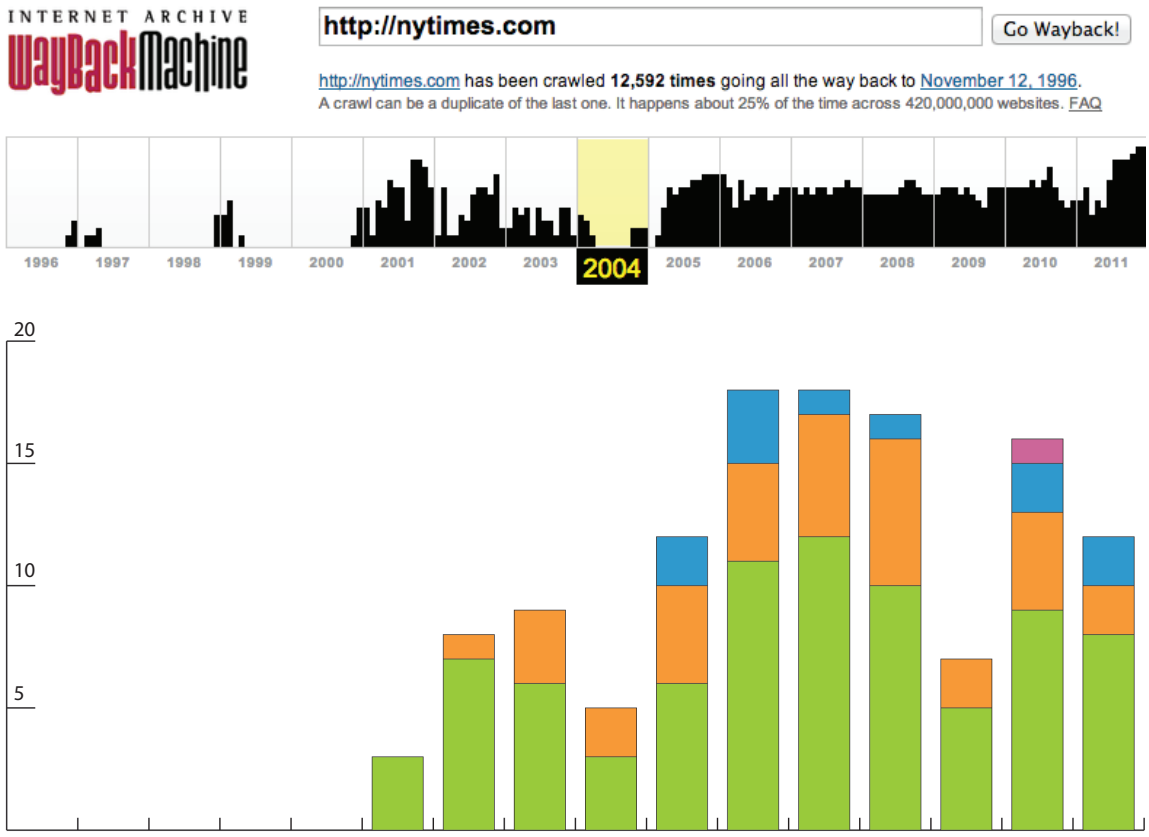

Figure 1: The number of trackers that have been detected by the Tracker Tracker tool per year on the archived NYT front pages in the IAWM between 1996 and 2011. The trackers have been color-coded according to the tracker types provided by Ghostery. Green: ad, orange: tracker, blue: analytics, pink: widget.

No third-party trackers have been detected between 1996 and 2000. As discussed previously, ad servers and trackers have been an integral part of the web since the mid 1990s, so I manually verified the data of this period by checking the source code for the presence of third-party trackers. In 1996 and 1997 the front page of the NYT is a clickable image map and does not contain any trackers. In 1998 the NYT contains a first-party tracker from RealMedia ads (now 24/7 Media) from the installed ad management platform Real Media on the NYT domain.[9] As of 2001 the NYT has started using a number of third-party

[9] First party trackers are issued from the same domain as the website "that the user has voluntarily interacted with" whilst third-party trackers are issued from a different domain than the website, indicating involuntary interactions with an external actor (Mayer and

advertising services: DoubleClick, LinkShare and Microsoft Atlas. The number of trackers increases per year and in 2006 and 2007 the NYT front page contains 18 unique trackers over that year. After 2007 there is a decline in the number of unique trackers which reflects the findings of the previously mentioned longitudal tracker study by Krishnamurthy and Wills who found an "increasing aggregation of user-related data by a steadily decreasing number of entities" (2009, p. 541). Further research could address this phenomenon by looking into whether this indicates media concentration. This would be a way to investigate how larger cultural, social and economic patterns on the web might be reflected in a website's ecosystem, such as increasing ownership concentration in the ad network industry.

Trackers in the form of widgets, which include social plugins such as Like, Share or Tweet buttons, are relatively absent from the results. This can be explained by the setup of the research design. In this study I focused on the front page of the NYT, whilst widgets such as social buttons are often not implemented on the front page but on the single-article page to Like, Share and Tweet an article. Figure 2 shows the diverse tracker environment the NYT website has been embedded in over the years.

Mitchell 2012, 413). In this chapter a tracker refers to a third-party tracker, indicating a connection with an external actor.

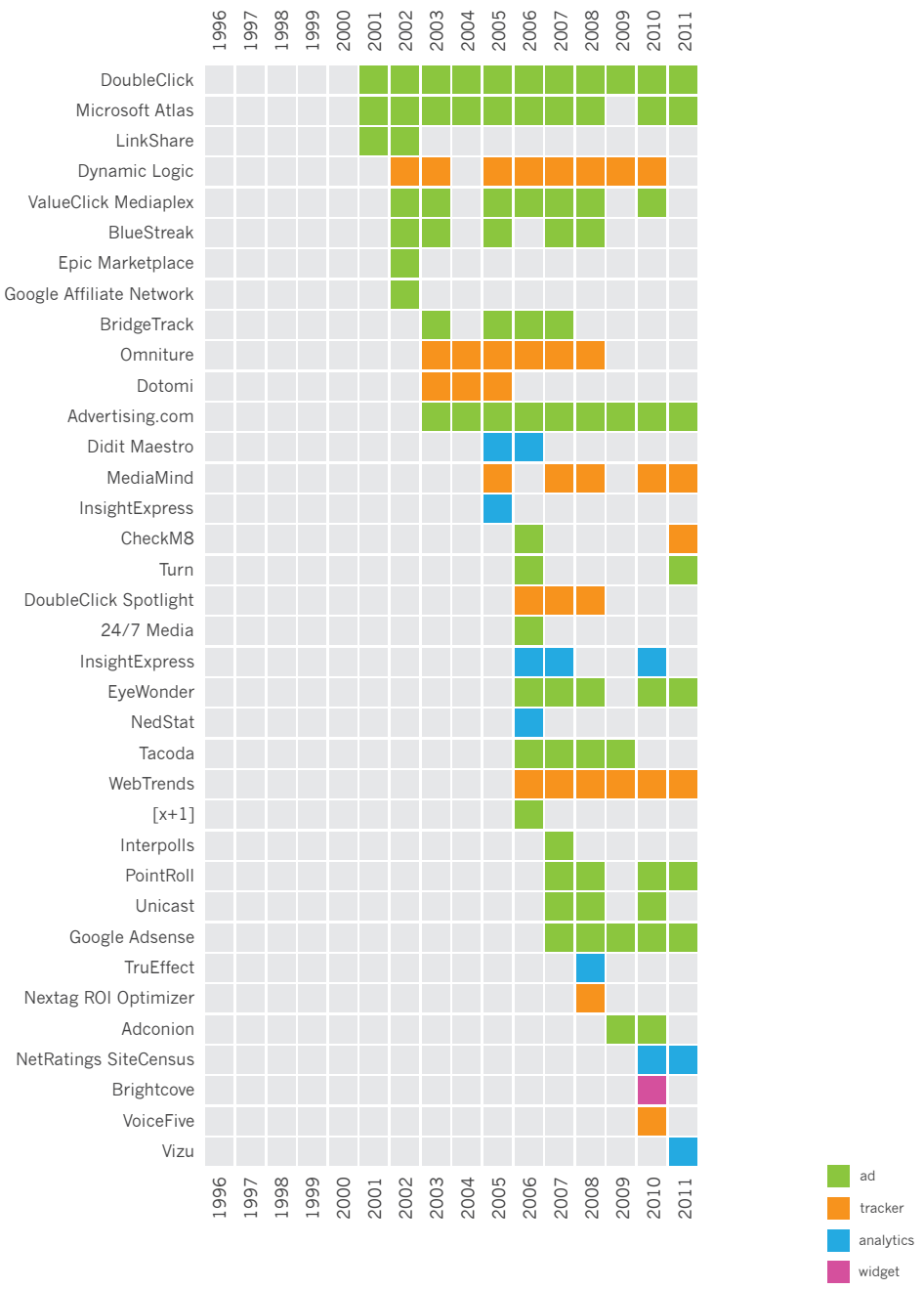

Figure 2: Names and types of the trackers that have been detected by the Tracker Tracker tool per year on the archived NYT front pages in the IAWM between 1996 and 2011. The trackers have been color-coded according to tracker type, see Figure 1.

This case study on the NYT has demonstrated how we can employ historical source code analysis to analyze the historical tracker environment of a website using the IAWM. As such it has put forward a way in which an individual archived website can be used to uncover the interactions between the website and its environment. One of the limitations of the case study is that the pattern of a specific tracker may change over time by using different scripts or tracking techniques (Orr, Chauhan, Gupta, Frisz, & Dunn, 2012). However, this case study has shown that in many cases patterns are established on domain name level and do not change significantly over time, e.g. [ad.doubleclick.com].

In future research the method presented in this chapter could be used to scan a larger set of websites and move beyond the front page of these websites to also detect social buttons. Scanning a large collection of archived websites over the timespan of 10 years would allow for reconstructing and analyzing the changing techno-commercial configurations of the web, for example by focusing on the changing infrastructural and economic underpinnings of the web with the rise of social media. This case study has demonstrated how a web historiographical approach may employ existing web archives to study a website's ecosystem over time in order to analyze historical states of the web using the website's archived source code.

Web histories: Reconstructing past states of the web using source code snippets

This chapter has aimed to contribute to the growing field of web historiography by putting the IAWM to new uses. One of the questions within this field is what kinds of web histories can be told using web archives. Dominant approaches focus on the history of a single site or a network of sites through historical hyperlink analysis because of the IAWM's focus on the unit of the website at the expense of the website's larger context (Ben-David & Huurdeman, 2014; Brügger, 2013; Rogers, 2013). I have shown how the code snippets of third-party objects in the archived source code can be used to address a common problem that web historians are facing: missing context. The archived source code contains information about a website's relationships with third parties and can provide an entry point for reconstructing a website's historical ecosystem. In this chapter I have developed and positioned *historical source code analysis* as a method to move beyond content analysis, single-site analysis, and hyperlink analysis of websites over time. Whilst the source code

is useful to analyze a single website's development or content over time, I have outlined an approach for using it to explore changes in the composition of the web more broadly. As I have argued, the presence of third-party objects, scripts, and tools within a collection of websites can be employed for doing web history. The use of analytics, social plugins, and advertising, whose tracking capacities can all be detected via the Tracker Tracker tool, provides an important starting point to analyze the business strategies and economic underpinnings of websites and the web at large. That is, historical source code analysis enables empirically investigating infrastructural and economic changes on the web. Furthermore, such studies can also be of value for other areas of research, for example privacy researchers who wish to examine how techniques for gathering user data and the types of user data via websites have changed over time. Software studies scholars,[10] as also proposed by Ankerson (2009), may employ historical source code analysis to understand the commercialization of the web and the practices of web masters to investigate the web's underlying infrastructure and how websites, users and other actors on the web are connected. These are important issues to consider for understanding the foundations of how the web is built, organized and supported.

To do so, in this chapter I have proposed to reconceptualize the study of websites as website ecology. By developing a method which enables web historians to scan websites for trackers and other third-party objects over time, I have propositioned to add the study of historical website ecosystems to the field of web historiography.

[10] My proposal for historical source code analysis can also be seen as a contribution to Critical Code Studies (CCS), an approach aligned with software studies and platform studies which analyzes the code layer of software (Marino, 2014). It specifically addresses a concern expressed by Matthew Kirschenbaum during the 2011 HASTAC Scholars Critical Code Studies Forum that "by focusing on the analysis of code snippets, CCS could potentially abstract code from its larger software constructs" (2014). The approach put forward here uses code to direct the attention back to these larger software constructs of the web such as tracking ecologies.

Acknowledgements

I would like to thank Richard Rogers, Anat Ben-David, Carolin Gerlitz, Fernando van der Vlist, the reviewers, and editor Niels Brügger, for their feedback on earlier drafts of this chapter.

References

Ammann, R. (2011). Reciprocity, Social Curation and the Emergence of Blogging: A Study in Community Formation. *Procedia - Social and Behavioral Sciences*, *22*, 26–36. http://doi.org/10.1016/j.sbspro.2011.07.053

Ankerson, M. S. (2009). Historicizing Web Design: Software. Style, and the Look of the Web. In J. Staiger & S. Hake (Eds.), *Convergence, Media, History* (pp. 192–203). New York, NY: Routledge.

Ankerson, M. S. (2012). Writing web histories with an eye on the analog past. *New Media & Society*, *14*(3), 384–400.

Bastian, M., Heymann, S., Jacomy, M., & others. (2009). Gephi: an open source software for exploring and manipulating networks. *ICWSM*, *8*, 361–362.

Ben-David, A. (2012). *Palestinian Border-making in Digital Spaces*. Bar Ilan University, Ramat Gan, Israel.

Ben-David, A., & Huurdeman, H. (2014). Web Archive Search as Research: Methodological and Theoretical Implications. *Alexandria*, *25*(1), 93–111.

Berry, D. M. (2012, September 23). Life in Code and Software. Retrieved April 21, 2013, from http://www.livingbooksaboutlife.org/books/Life_in_Code_and_Software/Introduction

Borra, E., Den Tex, E., Gerlitz, C., Helmond, A., Martens, K., Rieder, B., & Weltevrede, E. (2012). *Tracker Tracker*. Amsterdam, Netherlands: The Digital Methods Initiative. Retrieved from https://tools.digitalmethods.net/beta/trackerTracker/

Borra, E., Weltevrede, E., Helmond, A., Stevenson, M., De Vries Hoogerwerff, M., & Rogers, R. (2009). *Internet Archive Wayback Machine Link Ripper*. Amsterdam, Netherlands: The Digital Methods Initiative. Retrieved from

https://tools.digitalmethods.net/beta/internetArchiveWaybackMachineLinkRipper/

Brügger, N. (2008). The archived website and website philology: A new type of historical document? *Nordicom Review*, *29*(2).

Brügger, N. (2009). Website history and the website as an object of study. *New Media & Society*, *11*(1-2), 115–132.

Brügger, N. (2013). Web historiography and Internet Studies: Challenges and perspectives. *New Media & Society*, *15*(5), 752–764.

D'Angelo, F. (2009, October 26). Happy Birthday, Digital Advertising! Retrieved from http://adage.com/article/digitalnext/happy-birthday-digital-advertising/139964/

De Souza, C., Froehlich, J., & Dourish, P. (2005). Seeking the source: software source code as a social and technical artifact. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work* (pp. 197–206). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1099239

Elmer, G. (2002). The case of web browser cookies: enabling/disabling convenience and relevance on the Web. In *Critical Perspectives on the Internet*. Lanham, MD: Rowman & Littlefield.

Elmer, G. (2006). Re-tooling the Network Parsing the Links and Codes of the Web World. *Convergence: The International Journal of Research into New Media Technologies*, *12*(1), 9–19. http://doi.org/10.1177/1354856506061549

Elmer, G., & Langlois, G. (2013). Networked campaigns: Traffic tags and cross platform analysis on the web. *Information Polity*, *18*(1), 43–56. http://doi.org/10.3233/IP-2011-0244

Foot, K., & Schneider, S. (2010). Object-oriented web historiography. In N. Brügger (Ed.), *Web History* (pp. 61–79). New York: Peter Lang. Retrieved from http://faculty.washington.edu/kfoot/Publications/Foot_Schneider.pdf

Fuller, M. (2005). *Media Ecologies: Materialist Energies in Art and Technoculture*. Cambridge, MA: The MIT Press.

Gehl, R. (2014). *Reverse Engineering Social Media: Software, Culture, and Political Economy in New Media Capitalism*. Philadelphia: Temple University Press.

Gerlitz, C., & Helmond, A. (2013). The Like economy: Social buttons and the data-intensive web. *New Media & Society*, *15*(8), 1348–1365. http://doi.org/10.1177/1461444812472322

Ghostery. (2013, October 1). Ghostery™ FAQs. Retrieved January 10, 2013, from http://www.ghostery.com/faq

Goddard, M., & Parikka, J. (2011). Unnatural ecologies. *The Fibreculture Journal*, *17*. Retrieved from http://seventeen.fibreculturejournal.org/

Kirschenbaum, M. G. (2003). Virtuality and vrml: Software Studies after Manovich. *Electronic Book Review*, (The Politics of Information). Retrieved from http://www.electronicbookreview.com/thread/technocapitalism/morememory

Krishnamurthy, B., & Wills, C. (2009). Privacy diffusion on the web: a longitudinal perspective. In *Proceedings of the 18th international conference on World wide web* (pp. 541–550). New York, NY, USA: ACM. http://doi.org/10.1145/1526709.1526782

Latour, B. (1999). *Pandora's Hope: Essays on the Reality of Science Studies* (1 edition). Cambridge, MA: Harvard University Press.

Liu, A. (2004). Transcendental data: Toward a cultural history and aesthetics of the new encoded discourse. *Critical Inquiry*, *31*(1), 49–84.

Manovich, L. (2013). *Software Takes Command*. New York, NY: Bloomsbury Academic.

Marino, M. C. (2014). Field Report for Critical Code Studies 2014. *Computational Culture*, (4). Retrieved from http://computationalculture.net/article/field-report-for-critical-code-studies-2014

Mayer, J. R., & Mitchell, J. C. (2012). Third-party web tracking: Policy and technology. In *Security and Privacy (SP), 2012 IEEE Symposium on* (pp. 413–427). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6234427

Orr, C. R., Chauhan, A., Gupta, M., Frisz, C. J., & Dunn, C. W. (2012). An Approach for Identifying JavaScript-loaded Advertisements Through Static Program Analysis. In *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society* (pp. 1–12). New York, NY, USA: ACM. http://doi.org/10.1145/2381966.2381968

Park, H. W. (2003). Hyperlink network analysis: A new method for the study of social structure on the web. *Connections*, *25*, 49–61.

Postman, N. (1970). The Reformed English Curriculum. In A. C. Eurich (Ed.), *High School 1980: The Shape of the Future in American Secondary Education* (pp. 160–168). New York: Pitman.

Roesner, F., Kohno, T., & Wetherall, D. (2012). Detecting and Defending Against Third-party Tracking on the Web. In *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation* (pp. 12–12). Berkeley, CA, USA: USENIX Association. Retrieved from http://dl.acm.org/citation.cfm?id=2228298.2228315

Rogers, R. (2002). Operating issue networks on the Web. *Science as Culture*, *11*(2), 191–213.

Rogers, R. (2013). *Digital Methods*. Cambridge, MA: The MIT Press.

Samar, T., Huurdeman, H., Ben-David, A., Kamps, J., & de Vries, A. (2014). Uncovering the Unarchived Web. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1199–1202). New York, NY, USA: ACM. http://doi.org/10.1145/2600428.2609544

Schneider, S. M., & Foot, K. A. (2004). The web as an object of study. *New Media and Society*, *6*, 114–122.

Schneider, S. M., & Foot, K. A. (2010). Object Oriented Web Historiography. In *Web History*. New York: Peter Lang Publishing.

Song, F. W. (2010). Theorizing web 2.0: A cultural perspective. *Information, Communication & Society*, *13*(2), 249–275.

Stevenson, M. (2010). The archived blogosphere: exploring web historical methods using the Internet Archive. Presented at the Digital Methods mini-conference, University of Amsterdam.

Turow, J. (2011). *The Daily You: How the New Advertising Industry Is Defining Your Identity and Your Worth*. New Haven: Yale University Press.

van der Velden, L. (2014). The Third Party Diary: Tracking the trackers on Dutch governmental websites. *NECSUS. European Journal of Media Studies*, *3*(1), 195–217.

Weltevrede, E. (2009). *Thinking nationally with the web: A medium-specific approach to the national turn in web archiving*. University of Amsterdam, Amsterdam.

Weltevrede, E. (2016). *Repurposing digital methods: The research affordances of platforms and engines* (Ph.D.). University of Amsterdam, Amsterdam. Retrieved from https://wiki.digitalmethods.net/Dmi/RepurposingDigitalMethods

Weltevrede, E., & Helmond, A. (2012). Where do bloggers blog? Platform transitions within the historical Dutch blogosphere. *First Monday*, *17*(2). http://doi.org/10.5210/fm.v17i2.3775