R. Rogers (2018). Periodizing web archiving: Biographical, event-based, national and autobiographical traditions, in Niels Brügger and Ian Milligan (eds.), SAGE Handbook of Web History. London: Sage.

Periodizing web archiving: Biographical, event-based, national and autobiographical traditions

Richard Rogers

Abstract

Since the founding of the Internet Archive in mid-1990s, approaches to Web archiving have evolved from striving to save all websites to focusing efforts on those dedicated to riveting events (elections and disasters), national heritage and most recently the self in social media. Each approach implies or affords a certain historiography: site-biographical, event-based, national and autobiographical (or selfie) history writing. Having proposed a periodization of the history of web archiving and the kinds of histories implied by each period's dominant approach, the article turns to the so-called 'crisis' in scholarly web archiving use, and proposes a methodological imagination to address it. Among the digital methods put forward to repurpose existing web archives, one may make screencast documentaries about the history of the web, create thematic collections and query them for social history purposes, conjure a past state of the web through historical hyperlink analysis and discover missing materials, and finally examine websites' underlying code allowing for the study of tracking over time. In all the piece calls for inventive methods to invite the further use of web archives.

Keywords

Web archives, Internet Archive, Wayback Machine, digital methods, web history, media history, digital history

Introduction: Historiographies built into web archives

The purpose of this chapter is to periodize web archiving, in order to discuss four ongoing traditions that form an overlapping and layered history of both the implementation but more so the study and use of archived websites. The most contemporary period of web archiving, as with most recent periods generally, is perilous to characterize, but the self-archiving and selfie culture undertaken most visibly on social media platforms as Facebook and Instagram, appears to sit at the end of a spectrum that commenced with the Internet Archive and preserving single websites, and has witnessed a chronology of efforts that include both event-based as well as national web making. I argue that each of these periods corresponds to a particular historiographical tradition, inviting certain kinds of content-based history-writing with the web archives. The piece discusses not only the research that each tradition of web archiving affords but also approaches to studying web archives that (perhaps counter-intuitively) are not based on a study of website content. Code-based analysis of archived websites allows for avenues of research both about web archiving – such as the historical reconstruction of the websites missing in archives (through an historical hyperlink network analysis) – as well as histories of the web that are broader social observations such as the rise of tracking that gathers data on web users (through an overtime analysis of trackers and cookies embedded in archived websites). Ultimately the chapter concludes with a discussion of the so-called 'crisis in web archive use' and ways to address it, such as repurposing and building atop existing web archives.

To trace the history of web archiving, I argue, is to appreciate the distinctive historiographical points of view built into web archives since the mid-1990s: from the

biographical (or single site) and the event-based to the national and autobiographical traditions. A discussion of each would concern the implications for (web) history-writing as well as web archiving practices, including a call for the consideration of other scholarly uses of web archives. In order to address the 'crisis in web archive use', which primarily concerns their under-utilization, one may look to such creative uses as making screencast documentaries of website histories, curating thematic or issue-based collections from existing archives, undertaking historical hyperlink analysis to conjure a past state of the web as well as examining the underlying code of archived websites (for cookies, for example, in order to study tracking or surveillance). It also may be of interest to expand upon archival regimes that privilege only the national and 'official' definitions of the public interest that drive selection (national methodologies), in order to enable historical projects that are currently unintended, such as a reconstruction of the Yugoslav web (Ben-David, 2016).[1]

Before developing the argument, I would like to mention four caveats about the periodization. The first is that I offer a contemporary rather than a sweeping periodization of *longue durée* such as the medieval, early modern, modern or postmodern'. The time span is but two decades. The second is that periodization does not shutter one tradition as it calls another into ascendancy. Whilst epistemologically distinctive and related to specific touchstones in the history of web archiving, the biographical, event-based, national and autobiographical historiographies overlap in time. Each may endure, but one may lack the vibrancy of its early period, having experienced pioneer's regress; it may have matured. The third caveat is that the periodization does not rest upon a single cause of change, whereupon one would follow back-end technology, front-end interface, institutional involvement or academic paradigm formation (to name a few), and be able to identify the triggering event from one area that prompted the next period. Moreover, the effort here is not to

develop a framework that explains (social) change in web archiving outlooks, however much the four areas I just alluded to may be of assistance. Finally, the periodization itself is from a point in time, and as such could be rewritten anew in future, or itself be conceptualized as a period in theory, e.g., the early days of web archiving theory when initial thoughts were formed on how to periodise web archiving history.

The periodization rests on touchstones that in retrospect brought into being archival regimes or repertoires that inform the web historiographies on offer. Here the notion of archival regime refers to the work of Wolfgang Ernst, defining it as 'not an idiosyncratic choice, but a rule-governed, administratively-programmed operation of inclusions and exclusions' (2006: 114). These regimes inform the kinds of histories that are in some sense given by the archives, or afforded to be written, if you will. The archives are also historicizable, and may be studied for their periodicity, or reflecting a particular web time when they were created.

Single-site histories, or the biographical tradition

The Internet Archive was founded by Brewster Kahle in 1996 to archive 'everything': 'I usually work on projects from the "you've-got-to-be-crazy stage"' (Reiss, 1996; Livingston, 2008; Kahle & Parejo Vadillo, 2015). The Sisyphean improbability of the 'everything archive' project (or its 'craziness') lies in at least two elements: one may never capture all of the web (at any one time), and as it grows one is presumably increasingly capturing less of it. At its very outset the Internet Archive presented itself less in the service of future generations, web and other history scholars, the court of law (and copyright infringement) or other use cases to which it subsequently lent itself. Rather, it offered a solution to a major issue of its time, the '404 File Not Found' problem (Kahle, 1997). Beginning at least in February 1998, Alexa Internet,

shortened from Alexandria, made available a browser toolbar with a small button that would pulsate if the website visited on the live web was not found (1998). 'WayBack,' as the button was named, remarkably supplied the missing page from the Internet Archive.

In exchange for receiving these offline webpages from the archive, one would agree to allow one's surfing history to be logged so that the archive's crawlers could visit the sites to archive them. That was the data exchange, in one of the earlier 'free' business models (Anderson, 2009). Through the Alexa toolbar, one also could retrieve information about the website one was visiting, such as how fresh (or stale) it was. One also can view related websites, the speed with which the website loads, and even its inlink count as a sign of its popularity or authority. The Wayback Machine (and the toolbar) are thus period pieces, and as such of interest to web historians in how they capture the mid-1990s, where loading speeds, broken links and related destinations were all considered of the moment for the 'surfer'.

Indeed, even without the toolbar the Wayback Machine, launched in 2001, later maintained the web as a surfer's medium rather than the searcher's medium that came after it. The surfing continued within the archive itself. Rather than allowing broken pathways, the archive directed hyperlinks on archived webpages to take the user to the page closest in time to the original, or to the live webpage, if it had not been saved. This multi-temporal archival experience arguably made surfing into one of the Wayback Machine's uses, together with various forms of historical reconstruction.

From a historiographical point of view, the Internet Archive, with its Wayback Machine, organized the history of the web in a specific manner. Whilst since 2016 one may search for keywords (that appear in URLs or in metatext), the Wayback

Machine from the beginning invited one to type in a single URL, such as http://www.google.com, and navigating its history through a listing or calendar of archived instances as well as a timeline.[2] The interface thus sees the web as a history of single websites (or more specifically URLs) that evolve through time; one may view their stories through the interface.[3]

The default setting on the timeline is a monthly impression, with a chunky arrow that invites the user to click through the changes to the website at that historical pace, though through the calendar there is also the opportunity to choose specific dates as well as times of the day of those frequently archived sites. The ones archived most frequently are likely from 'focused crawls' rather than the 'broad crawls' that capture the vast majority of distinct sites in the archive (Kimpton et al., 2003; Sigurðsson, 2005).

Event-based special collections, or event-based historiography

The second period's touchstone is the Webarchivist project, and its readiness when 9/11 transpired (2001). There were collection concepts ('web sphere'), server capacity and researchers in place when the airplanes struck the World Trade Center and the Pentagon, out of which came not only the pioneering 9/11 web collection but solidified event-based web archiving, an *idée fixe* in the selection repertoire of web archives. The historiographical tradition it spawned is thus based on collecting websites around events, predominantly elections, disasters and (papal and presidential) transitions. It actually began with the Internet Archive's first project (with the Smithsonian), archiving the 1996 U.S. presidential elections, and the Webarchivist project followed in those footsteps with its plans to capture the 2002 U.S. congressional and other elections, when disaster struck on September 11, 2001. With this agile, just-in-time archiving, they were able to amass some 25,000 websites

(over some four months), now part of the U.S. Library of Congress's 'digital collections'. Subsequently, the Webarchivist project continued to collect websites around elections and further expanded its disaster collection work with the special collection surrounding the Asian Tsunami of 2004 (1,500 websites).

Event-based archiving of 'elections and disasters' owe greatly not only to the Webarchivist project but also to a particular pioneering collection technique. 'Web sphere analysis' is a demarcation technique that defines a collection space substantively, temporally as well as web-topologically. Foot and Schneider describe a web sphere as 'a set of dynamically defined digital resources, often connected by hyperlinks, spanning multiple websites relevant to a central event, concept or theme and bounded temporally' (Foot & Schneider, 2002: 225; Schneider & Foot, 2005). Akin to Kahle's 'everything' collection which necessarily evolves as it discovers new websites, the web sphere definition is also innovative for its webbiness, allowing for its dynamic collection through discovering new websites through hyperlink analysis. It is also historicizable, having borrowed from the emergence at the time of hyperlink mapping techniques to map networks of websites around the same issue (Rogers, 2012). The project also was contiguous with contemporaneous assumptions of the metaphysics of the web, portions of which are organized (and readily archivable) as spheres, like (in its day) the up-and-coming blogosphere.

National web archives, or national historiography

The third derives not so much a moment but rather a long march of the national institutions onto the web, and the growing acceptance of the very idea of the 'Danish part of the internet,' or the Swedish part or the Czech part, to be archived separately, either by law or custom (Schostag & Fønss-Jørgensen, 2012). Rather than a cyberspace (where everything can be collected) or a sphere (where a thematic event is

to be located), from the national point of view the web is carved into relatively tidy, geographical (content) portions for national institutions keeping public records as well as national heritage. In such an archival regime it would be remarkable for national institutions to collect subaltern materials that archive more than national content, be it (in order of importance) from the top-level country domain, lists of nationally significant websites and social media pages, events as well as language detection.

The third corresponding historiographical tradition is thus the national, which began (as with the event-based tradition) as an Internet Archive service to Swedish and Icelandic national libraries, before they and other national libraries settled into their own collecting in the mid-2000s. The Internet Archive would crawl a top-level country domain and make it available to the national library for preservation and access, whereas nowadays, in a turn of the tables, national libraries crawl their 'own' domains, without subsequently uploading them as collections to the Internet Archive (often out of concern for copyright infringement, which also keeps them offline and accessible only on site), with the exception of the Portuguese which both uploads and makes accessible its archive online (Internet Archive, 2017). Indeed, one of the stated aims of the national web archiving projects is to reduce dependence on foreign services (Gomes et al., 2008). Most other national web archiving initiatives gradually have developed their own archiving capacity, with particular definitions of what constitutes the national.

National libraries and archives are tasked with archiving public records as well as other content of national interest, so the question asked with regards to web archiving by national libraries has to do with what constitutes public interest. What should count, methodologically, as valuable 'national content'? Archiving traditions are often adjoined to the technical infrastructure of the web, so the national becomes the

websites using the top-level country domain as well as the language, together with public interest and heritage definitions. Along these lines the national libraries have existing appraisal and selection traditions as well as vehicles to undertake this kind of work, such as national deposit laws, which obligate archiving. Some countries, such as the Netherlands, do not have such deposit laws, but have similar principles concerning collecting and preserving the public record as well as cultural heritage. Whether with or without deposit laws, one could argue the collection approach preserves some combination of public record for official history and heritage for national history.

Among the influential definitions of the constitution of heritage is the Danish, which collects and stores 'Danica' or national cultural heritage, but also incorporates the history of web archiving into its regime, collecting a couple of events per year. For the Danes, as well as for other national libraries, websites that use the top-level country domain are to be archived as well as those intended for a Danish audience, and written in Danish. Websites about the Danish people, significant Danish personalities or well-known figures, or simply about Denmark (no matter the language) also would be candidates for archiving. Therefore, if there is a website in English about the 19th century Danish writer Hans Christian Andersen, that website could be included. In the crawling regimes, apart from regularly archiving significant websites and periodically archiving lesser ones, there are also 2-3 'events' (such as Danish elections) that Netarkivet, the archiving authority, is prepared to save. The Dutch (without the legal deposit) make use of a similar definition of 'the national' as do the Portuguese and French, whilst the latter is somewhat broader in its special collections – apart from the French elections, examples of special collections are more expansive than events, such as 'blogs, sustainable development, Web activism' (BnF,

2017). In all the descriptions of collection-making by national libraries, the national interest takes precedence.

Such an archiving routine editorializes the web as a national story, and moves web archivalism far away from the Internet Archive's initial approach of crowdsourcing URLs through people installing a toolbar in a 'grab them all' tradition, or the hyperlink analysis for event-based, 'web sphere' collections. It often turns 'events' largely into national ones. These days an Asian tsunami likely would not be archived by a European national library. As a case in point the last international event collected by the U.K. web archive is from 2004; the Danish archive appears increasingly to concentrate on national events, with the exception of international collaborations with for example the Czech archive and its Vaclav Havel collection.[4]

Autobiographical archiving

Finally, with the rise of social media (especially Facebook) and mobile platforms challenging the web as the dominant online content and activity space, has come the lessened capacity to archive it, or as archival studies scholars have put it: 'the responsibility of archiving Facebook data [lies with] individual users' (Sinn & Syn, 2013); hence the notion of the autobiographical. In practice, each of the three traditions mentioned above have the means to retain some social media (and occasionally do archive Facebook 'pages' such as We are all Khaled Said, crucial in the Egyptian Revolution of 2011). That archiving, however, does not pertain to individual profiles after login.

This fourth tradition (loosely defined) is the development of the capacity of archiving oneself, especially one's social media use, which is outside the purview of the other

web archiving traditions. One could refer to such activities as life blogging and quantifying oneself as precursors to self-archiving online, since these, often health-related authoring practices are often in the style of public diaries, worthy of saving like eighteenth-century chapbooks and other 'popular' sources of everyday life (Darnton, 2009). 'Mommy' as well as 'DIY fatherhood' blogging are examples from the web, whilst the 'wounded healers of Instagram' from social media (Ammari et al., 2017; Sanchez-Querubin, 2017). Posting entries on Facebook (and even Snapchat) could be construed similarly, however it is the request for the data dump that could be understood at least initially as self-archiving (Facebook, 2017). That is to say, awareness of the capacity to create collections of one's own history first relied on knowing one's digital rights (so to speak) and asking companies to comply with them. There have arguably been larger developments in terms of opportunities for self-archiving, and I would like to discuss four recent ones.

The first development concerns health apps, and the data retained from fitness bands and watches such as Apple's iWatch. Here there are opportunities for 'personal data requests' that relate less to one's writings than to one's physical state. The second development concerns the movement of activity from the web to smartphones, and the question as to what to archive. Apart from making software App collections (practiced by the Internet Archive, for example), there is a shift in collection from 'content' to data. One may access mobile-related activity through an API (Littman et al., 2017). Such is a more general strategy for social media archiving, be it originating from the web or from a mobile device (Thomson, 2016). The third development, recording a period of one's life in social media, is exemplified by the social media artistic 'performance' by Amalia Ulman called 'Excellences & Perfections' (2014). In a period of six months Ulman documents on Instagram her aspiration to become a Los Angeles 'it girl', gaining followers as she moves through aspirational stages of

becoming from ‘trying to be discovered’ to ‘bad girl’ and finally to ‘healthy lifestyle’ before finally (and surprisingly) posting ‘the end’. Rhizome, the digital arts group, in producing this work of art, introduced its Web Enact software to capture such coming of age or other personal developments on social media. Later called webrecorder.io, it allows one to ‘record’ a social media page. This approach stands in contrast to requesting a data dump, downloading or taking screenshots of the website, or tapping into the API for data. Finally, I would like to briefly mention ‘selfie archiving’, however much it has been undertaken by neither self-archivers nor web archivists. Researchers have created collections of Instagram photos hashtagged selfie. In one project by TIME magazine, the goal was to crown the ‘Selfiest Cities in the World’ (Wilson, 2014). The researchers created a database of 400,000 Instagram photos that were tagged #selfie, determined each selfie’s geolocation and created a ranked list of cities, where in the event Makati City and Pasig in the Philippines had the most selfies per 100,000 people, followed by Manhattan, Miami, and so forth; the list is dominated by Asian and North American cities that were subsequently plotted onto a map showing selfie density. Conducted by a team of researchers led by Lev Manovich, the second selfie archiving project – Selfiecity – created selfie collections from New York, São Paulo, Berlin, Bangkok and Moscow (using #selfie and city geolocations in the queries). The researchers subsequently measured the formal properties of the faces (tilt of the head, smile, etc.). Among other findings, the project found that selfies are largely an undertaking for 23 to 26-year-olds, though in Moscow they tend to trend older and in São Paulo younger. Furthermore, in Moscow people are slightly gloomier, whereas in São Paulo happier. Thus the project outputs city mood gradations. It should be noted that Instagram (and the moody selfies) are accessible on the web, though most users are mobile-based (on smartphones), pointing up a number

of issues such as the movement of users and their content production from the web to the App space and also how web archiving may assist with mobile content archiving.

What are the prospects for archiving social media institutionally? With respect to the single-site or biographical tradition (and the contents available via the Wayback Machine), attempts to crawl and archive social media as websites are fraught with issues. Whilst hardly clear-cut, one could put forward a distinction between social media that are principally social networking sites or user-generated content platforms, and expect the content platforms to be more open to crawling and archiving than the social networking ones (Thomson, 2016) (see Table 1). In the event, Facebook and LinkedIn (seeing themselves as principally social networking sites) do not allow archive crawlers, having opted out through placing a robots.txt file respected (and archived) by the Internet Archive; most other platforms allow the archiving of their 'about' pages, terms of service, FAQs and other non-password-projected webpages.[5] With respect to archiving the user-generated content, platforms (Flickr, Pinterest, Reddit, YouTube) allow archiving of its materials, though in certain cases the materials are personalized (and ranked), so there is skewing.

The overall exception is Twitter, which donates its historical tweets (and archive) to the Library of Congress, where they are stored. The girth of such big data, however, has proven too great to make it available to researchers (Zimmer, 2015). Tweet collections also made be made (but not shared with others) with software tools, such as the Digital Method Initiative's TCAT (Twitter Capture and Analysis Tool), which queries the streaming and REST APIs for hashtags, keywords and user accounts (as well as @mentions). U.S. President Donald Trump's tweet collection (i.e., just his user account's tweets) is significant as are collections made from the U.S. presidential elections of Hillary Clinton and Trump supporters, respectively. There are also

hashtag collections, such as the infamous #iranelection (around which a great deal of debate ensued concerning the very idea of a 'Twitter revolution'). Given rate limits questions of collection completeness arise; there are Twitter-owned services (GNIP) that provide historical tweet sets at often exorbitant prices.

[Table 1 about here]

Each of the other traditions sketched above would tackle archiving social media somewhat differently. From the event-based tradition, archiving of Facebook pages (rather than individual profiles) has been practiced, especially in the case of the Egyptian Revolution of 2011 (Runyon & Houlihan, 2012; Urgola & Runyon, 2016). In the national tradition, one may wish to archive national public figures and official institutions (Facebook pages as well as Twitter feeds), as is practiced by the UK National Archive, for example. As mentioned there is also a movement afoot to shift attention from the html to the data, and tap into and archive the API streams offered by the social media companies. One could imagine, methodologically, that there would be 'everything', 'event-based' as well as 'national' strategies for API polling. At least with Facebook, which closed down the availability of extracting personal information, including one's own as well as that of friends with its 1.2 API version (mid-2015), the full autobiographical (all data about oneself across Facebook) still would rely on the data dump, though Archive-It's software would allow one to grab one's own profile page, and webrecorder.io affords the means to capture its dynamic history (Rieder, 2015).

Addressing the crisis in web archive use

In 2010 web archiving theorists referred to a crisis in the scholarly use of web archives, for there were so few users (Dougherty et al., 2010). The crisis has been illustrated through regular queries in Google Scholar and Google Web Search for the use of the citation preferred by the Library of Congress; for some years now, typing 'Archived in the Library of Congress Web' into Google Scholar has returned very few results (Rogers, 2013). The search engine returns are mainly self-citations. The Library of Congress Web has largely special collections in the event-based tradition, which could account for the low usage. In the single-site tradition, the employment of the Wayback Machine of the Internet Archive in scholarly work is greater, however much most articles are about Wayback URLs as references in legal cases and academic papers (and how Wayback combats live link rot), rather than as sources for web or digital history, with some notable exceptions (Musso & Merletti, 2016; Chakraborty & Nanni, 2017). In the national tradition, most national libraries have very few visitors to their web archives, given access policies. The computers allocated to web archive use by the National Library of France often lie idle. Annual users of the web archives in Denmark and the Netherlands number perhaps in the double digits.[6]

How might one address the crisis? Are there signs that it may be abating? Scholars have introduced ways and means to increase the use of web archives, such as making available full text search, building tools to visualize the contents of web archives, treating web archives as big data, concatenating archives with a common interface, or promulgating the creation of one's own archives (Huurdeman et al. 2013; Padia et al., 2012; Hockx-Yu, 2011; Meyer et al., 2017; Cowls, 2017; Memento, 2015; Archive-IT, 2017). For (contemporary) historians working with (web) archivists, one may wish to make one's own archive with the Archive-It service, and study it, as in the case of the Egyptian Revolution of 2011 (the uprising of the 25th of January), a project of the

American University of Cairo. The historical account of the rise and fall of an Islamic punk scene is also illustrative of the approach of making a collection to study it (Dougherty, 2017).

Another general approach, described below, is making creative use of existing archives with digital methods (Rogers, 2009). By a digital methods approach is meant the repurposing of dominant devices and platforms for social and medium research (or, in this case, digital and web history). How to repurpose the Wayback Machine's output of single websites? One captures a website's history and plays it back in the style of time-lapse photography as a screencast documentary with voiceover (Rogers, 2017).[7] One example is the history of Google.com, called 'Google and the Politics of Tabs' (Rogers & Govcom.org, 2008). Using the Wayback Machine, all unique historical instances of google.com are captured, i.e., those pages that contained changes as signified by an asterisk on the Wayback Machine's (classic) user interface. These are loaded into a moviemaker (QuickTime), and then played back to choose themes for a voiceover. The story is told by noting the gradual changes made over time to the front page of google.com, especially to the tabs above the search box, where different services came and went, such as the Google directory, or the human-edited listing. After appearing with great fanfare in 2000 it was gradually relegated to the 'more' button, and then placed behind 'even more', until the directory was finally removed all together in 2004. Other services made it all the way to the front-page interface before disappearing, or have staying power, such as images (Google Image Search). The story told is a web history: in the screencast documentary of Google.com, there is an evident, gradual decline of the directory over time (and with it the human librarian), and the simultaneous rise of the algorithm or the back end of search. Other examples of screencast documentaries have been made for media

history (e.g., the evolution of a newspaper on the web) as well as digital history (the evolution of the U.S. White House, whitehouse.gov).

The second contribution to web archive usage is to create one's own thematic collection of websites from one or more existing archives, first demonstrated by the Dutch newspaper, the *NRC Handelsblad* (Dohmen, 2007). By largely using the Wayback Machine, the researchers made a collection of Dutch right-wing and right-wing extremist websites, and devised a keyword query strategy to answer the question: Is Dutch culture hardening and becoming more extreme? To answer this question, they examined how the language on the websites changed over time, demonstrating that on the right-wing websites in particular, the language used was becoming increasingly extreme, approximately that on the right-wing extremist sites. The researchers cautiously concluded that Dutch culture is hardening, thereby making a contribution, however modest, to social history. Here it should be noted that such a project could not be realized with a national archival regime that saves 'significant' websites in the public interest for heritage or, for that matter, an event-based approach. Being able to make a collection of extremist sites would demand an 'everything' approach, augmented perhaps by combing multiple national web archives that sometimes stray from the national (heritage) methodology.

The third example is a project that again creates a thematic collection of websites within existing web archives, and conjures a past state of the web, so as to study the missing web and its significance through historical hyperlink analysis. The past state in question is the early blogosphere. There is an archived website called Eaton Web, which for many years had an authoritative list of blogs, and it was considered the portal (or leading directory) for the blogosphere (albeit largely American and English-language ones). In the Summer of 2001, the website owner, Eaton, gave up listing

new blogs because of abundance. To the researchers this was the sign that the period of the 'early blogosphere' was over. Eaton's last list was batch queried in the Wayback Machine, and it was found that only 20% or so of blogs were missing (see Figure 1). Consequently, historical hyperlink analysis was performed, which enabled the researchers to determine the significance of each of the websites in the blogosphere, including the missing ones, according to network measures. Thus all the sites were given historical context, and the past 'blogosphere' was depicted (or conjured) as a network. Using a similar technique of historical hyperlink analysis, the evolution of the Dutch blogosphere also has been mapped (Weltevrede & Helmond, 2012).

[Figure 1 about here]

Finally, the fourth approach is to study not the content of a single website, a thematic set over time, or a past state of the web through hyperlink analysis (including the unarchived sites), but rather the underlying code of one or more websites. This is a technique colleagues and I discovered coincidentally by using the browser add-on Ghostery whilst visiting historical websites in the Wayback Machine. Ghostery shows third party elements embedded in web pages when visited, including trackers, third party cookies and so forth. With Ghostery enabled while visiting an archived web page, one can view historical trackers (van der Velden, 2014; Deville and van der Velden, 2015). These trackers may be captured over time for specific websites, showing for example the history of tracking behavior by the *New York Times* (see Figure 2).

[Figure 2 about here]

Conclusions: From single site histories to historical network analysis

The effort here is to periodize web archiving, in order to describe how archival regimes and periodicity produce distinctive historiographical traditions. Historicized and described are the single-site (or website biography) approach built into the interface of the Internet Archive (mid-1990s), the event-based tradition from the Webarchivist project (and particularly the 9/11 archive), the long march of the national institutions and heritage methodologies making their own national webs (mid-2000s), and a variety of efforts to save social media, and particularly individual accounts through recording (as well as data dumps) in the autobiographical tradition (early 2010s). There is a distinctive chronology of web archiving with particular key moments per tradition, such as the Alexa toolbar, the 'web sphere', the 'Danish part of the web', and webrecorder.io, but each archiving regime and associated historiographical tradition continues.

The dominant approaches to web archiving and their in-built historiographies do not preclude others, however much (with the exception of the Internet Archive and some smaller projects) the emphasis is increasingly on the public interest remit of a heritage institution.[8] As a result, one is able to write national (web) histories from the contents stored by the national institutions. Such an emphasis, however, would preclude projects as the reconstruction and transformation of the Iraqi web before, during and after the Iraq War, 2003-2011. Given the refrain concerning the urgency of archiving the web and born-digital content, a rejoinder would add, when does it become urgent to archive other cultures than one's own (Beunen & Schiphof, 2006)?

Web archives are also underutilized. Many only can be accessed from a library reading room; for the German national web archive, for example, one would sit

behind a national library computer in Frankfurt or Leipzig. Few do, creating an air of crisis in web archive use. Brightening the archive is the task of scholars researching archive use, and building tools atop. Other engagement comes from making one's own (with Archive-It) or curating collections from existing archives. Indeed, among the creative use of archives for scholarly purposes is to build a thematic collection (e.g., right-wing and right-wing extremist websites in the Netherlands) and pose research questions about (the hardening of) culture. Other uses of web archives put forward in the digital methods approach above include the technique of creating a screencast documentary of the history of a website, conjuring a past state of the web through historical hyperlink analysis and examining the underlying code of archived websites and fishing out trackers and third party cookies with the aid of a tracker database, in order to put forward a history of (online) surveillance or behavioral targeting. Each approach rests on the capacity to build software on top of an online web archive, query it, extract data and make derivative works, which is a research practice different from visiting a library's reading room and browsing its web archive.

Notes

1. National web archival regimes have taken root in at least the following countries, which are among the members of the International Internet Preservation Consortium (IIPC): Canada, Chile, China, Croatia, Czech Republic, Denmark, Estonia, Germany, Finland, France, Iceland, Ireland, Israel, Japan, Latvia, Luxembourg, the Netherlands, New Zealand, Norway, Poland, Portugal, Scotland, Serbia, Singapore, Slovenia, South Korea, Spain, Sweden, Switzerland, United Kingdom and the United States. Numbering some 50, there are additional representatives from provinces (Catalonia and Quebec) as well as ones from universities and other (governmental) institutions. See Niu, 2012.
2. The Memento Project (mementoweb.org), the initiative that strives to source and output archived websites from multiple web archives, also invites the single URL as input.

3. The Internet Archive originally was provided to users of the Alexa Toolbar as a solution to the 404 file not found error, providing single Web pages. The Wayback Machine, as its interface, also provides single web pages. Since one is prompted to type a URL into the interface, users presumably would type front pages, or website domains, into the interface, and proceed from there. I thus refer to the Wayback Machine as organizing single-site histories. There are other use cases, too, such as pasting a longer URL, in order to look up copyright infringement. One may paste a specific historical URL that is now offline, thereby employing it in the manner of the Alexa Toolbar of old.
4. When collaborating with other international institutions (such as with the Czech web archive) the Danish web archive occasionally makes collections that are not primarily or partially Danish in focus. Since the early days, however, there appears to be an increasing tendency to follow the Danish heritage preservation policy, and archive events where Denmark is involved (Olympics, EU elections) as well as particularly national events such as the teacher lock-out, a Copenhagen science festival and a national scandal involving credit card transaction monitoring (Netarkivet, 2017).
5. Web archiving software, such as Archive-It, has the option to ignore robots.txt, thereby enabling the archiving of social media pages, groups and profiles, if one is logged in, and has given user credentials. Otherwise one archives default prompt pages, such as Facebook's 'log in or sign up'. (Archive-It does have a default user, Charlie Archivist, who has no friends.) Archive-It's instructions are explicit in their call to exclude personal profiles (Lohndorf, 2017). It should be added that social media companies such as Facebook and LinkedIn explicitly prohibit crawling without permission, and list on the robots.txt exclusion page which types of Facebook content archive crawlers may access, i.e., Facebook pages (of public figures) rather than personal profiles).
6. In the Netherlands it often has been a handful of users. In Denmark special workshops at Netarkivet organized by the Netlab at Aarhus University have bolstered the numbers.
7. One employs the Wayback Machine link ripper to harvest the URLs of all historical instances of a webpage, from which one chooses the ones to screenshot. Loading a select list of URLs into 'Grab them all' or another batch screenshot generator creates files to be imported into a moviemaker.
8. The Common Crawl project (commoncrawl.org) is another exception.

References

Alexa (1998) ‘Support’, Alexa.com, https://web.archive.org/web/19980209020820/http://www.alexa.com:80/support/details/index.html

Ammari, Tawfiq, Schoenebeck, Sarita and Lindtner, Silvia (2017) ‘The Crafting of DIY Fatherhood’, Proceedings of CSCW '17, New York: ACM, 1109-1122.

Anderson, Chis (2009) *Free: The Future of a Radical Price*. New York: Hyperion.

Archive-It (2017) ‘Archive-It. Web archiving services for archives and libraries’, San Francisco: Internet Archive.

Ben-David, Anat (2016) ‘What does the Web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain’, *New Media & Society*, 18(7): 1103–1119.

Beunen, Annemarie and Schiphof, Tjeerd (2006) ‘Legal aspects of web archiving from a Dutch perspective’, Centre for Law in the Information Society. Leiden: University of Leiden.

BnF (2017) Digital legal deposit: four questions about Web Archiving at the BnF, Paris: Bibilotheque national de France, http://www.bnf.fr/en/professionals/digital_legal_deposit/a.digital_legal_deposit_web_archiving.html.

Chakraborty, Anwehst and Nanni, Federico (2017) 'The changing faces of science museums: A diachronic analysis of museum websites', in Niels Bruegger (ed.), *Web 25: Histories of the First 25 Years of the World Wide Web*. New York: Peter Lang.

Darnton, Robert (2009) The Case for Books: Past, Present and Future. New York: Public Affairs.

Deville, Joe and van der Velden, Lonneke (2015) 'Seeing the Invisible Algorithm', in Louise Amoore and Volha Piotukh (eds.), *Algorithmic Life: Calculative Devices in the Age of Big Data*. London: Routledge. pp.87-105.

Dohmen, Jaap (2007) 'Opkomst en ondergang van extreemrechtse sites', *NRC Handelsblad*, 25 August.

Dougherty, Meghan, Meyer, Eric T., Madsen, Christine, van den Heuvel, Charles, Thomas, Arthur and Wyatt, Sally (2010) 'Researcher Engagement with Web Archives: State of the Art', London: JISC.

Ernst, Wolfgang (2006) 'Dis/continuities: Does the Archive Become Metaphorical in Multi-Media Space?' in Wendy Chun and Thomas Keenan (eds.), *New Media, Old Media. A History and Theory Reader*. New York: Routledge. pp.105-123.

Foot, Kirsten A. and Schneider, Steven M. (2002) 'Online Action in Campaign 2000: An Exploratory Analysis of the U.S. Political Web Sphere', *Journal of Broadcasting & Electronic Media*, 46(2): 222-244.

Gomes, Daniel, Nogueira, André, Miranda, João and Costa, Miguel (2008) 'Introducing the Portuguese web archive initiative', Proceedings of IWAW '08, Heidelberg: Springer.

Hockx-Yu, Helen (2011) 'The Past Issue of the Web', Proceedings of WebSci '11, New York: ACM.

Memento (2015) Memento Guide - Introduction to Memento. Mementoweb.org.

Huurdeman, Hugo C., Ben-David, Anat and Samar, Thaer (2013) 'Sprint methods for web archive research', Proceedings of WebSci '13, New York: ACM.

Internet Archive (2017) Arquivo.pt: the Portuguese web-archive, San Francisco: Internet Archive, https://archive.org/details/portuguese-web-archive.

Kahle, Brewster (1997) 'Archiving the Internet', *Scientific American*, March.

Kahle, Brewster and Parejo Vadillo, Ana (2015) 'The Internet Archive: An Interview with Brewster Kahle', *19: Interdisciplinary Studies in the Long Nineteenth Century*, issue 21.

Kimpton, Michele, Stata, Raymie and Mohr, Gordon (2003) 'Internet Archive Crawler Requirements Analysis', Heritix Internet Archive Webteam Confluence. San Francisco: Internet Archive.

Littman, Justin, Chudnov, Daniel, Kerchner, Daniel, Peterson, Christie, Tan,

Yecheng, Trent, Rachel, Vij, Rajat and Wrubel, Laura (2016) 'API-based social media collecting as a form of web archiving', *International Journal on Digital Libraries*, 28: 1-18.

Livingston, Jessica (2008) Founders at Work: Stories of Startups' Early Days. New York: Apress.

Lohndorf, Jillian (2017) 'Archiving Facebook', Archive-It Help Center, September, https://support.archive-it.org/hc/en-us/articles/208333113-Archiving-Facebook.

Musso, Marta and Merletti, Francesco (2016) 'This is the future: A reconstruction of the UK business web space (1996–2001)', *New Media & Society*, 18(7): 1120–1142.

Niu, Jinfang (2012) 'An Overview of Web Archiving', *D-Lib Magazine*, 18(3/4).

Padia, Kalpesh, AlNoamany, Yasmin and Weigle, Michele C. (2012) 'Visualizing Digital Collections at Archive-It', Proceedings of JCDL'12, New York: ACM.

Reiss, Spencer 1996. 'Internet in a Box', *Wired*, 1 October.

Rieder, Bernhard (2015) 'The end of Netvizz (?)', blog post, The Politics of Systems blog, http://thepoliticsofsystems.net/2015/01/the-end-of-netvizz/.

Rogers, Richard (2009) The End of the Virtual: Digital Methods. Amsterdam: Amsterdam University Press.

Rogers, Richard (2012) ‘Mapping and the Politics of the Web’, *Theory, Culture & Society*. 29(4/5): 193-219.

Rogers, Richard (2013) Digital Methods. Cambridge, MA: MIT Press.

Rogers, Richard (2017) Doing Web History with the Internet Archive: Screencast Documentaries. *Internet Histories*, 1(1/2): 160-172.

Rogers, Richard and Govcom.org (2008) ‘Google and the Politics of Tabs’, video, Amsterdam: Govcom.org.

Runyon, Carolyn and Houlihan, Meggan (2012) ‘Revolutionary Libraries: Building Collections and Promoting Research about the January 25th Uprising in Egypt’, *Alexandria*, 23(2): 73-77.

Sanchez-Querubin, Natalia (2017) ‘The Wounded Healers of Instagram’, Paper presented at Trauma Studies in the Digital Age workshop, Netherlands Institute for Advanced Study in the Humanities and Social Sciences, Amsterdam, 10-12 May.

Schneider, Steven M. and Foot, Kirsten A. (2005) ‘Web sphere analysis: an approach to studying online action’, in Christine Hine (ed.), *Virtual Methods: Issues in Social Research on the Internet*. Oxford: Berg. pp.157-170.

Schostag, Sabine and Fønss-Jørgensen, Eva (2012) ‘Webarchiving: Legal Deposit of Internet in Denmark. A Curatorial Perspective’, *Microform & Digitization Review*, 41(3/4): 110–120.

Sigurðsson, Kristinn (2005) ‘Incremental crawling with Heritrix’, Proceedings of IWAW ‘05, Vienna.

Sinn, Donghee and Syn, Sue Yeon (2014) ‘Personal documentation on a social network site: Facebook, a collection of moments from your life?’, *Archival Science*, 14(2): 95–124.

Thomson, Sara Day (2016) ‘Preserving Social Media: DPC Technology Watch Report 16-01’, Glasgow: Digital Preservation Coalition.

Ulman, Amalia (2014) ‘Excellences and Perfections’, New York: Rhizome, http://webenact.rhizome.org/excellences-and-perfections.

Urgola, Stephen and Runyon, Carolyn (2016) ‘Participatory Archives: Building on Traditions of Collaboration, Openness, and Accessibility at the American University in Cairo’ in Raymond Pun, Scott Collard and Justin Parrott (eds.), *Bridging Worlds: Emerging Models and Practices of U.S. Academic Libraries Around the Globe*. Chicago: Association of College and Research Libraries. pp.91-103.

van der Velden, Lonneke (2014) ‘The Third Party Diary: Tracking the trackers on Dutch governmental websites’, *NECSUS. European Journal of Media Studies*, 25 June.

Webarchivist (2001) ‘Please help us build a Web Archive of the Sept 11 Attack’, webpage, *Webarchivist.org*, 1 October, https://web.archive.org/web/20011001200536/http://webarchivist.org:80/.

Weltevrede, Esther and Helmond, Anne (2012) ‘Where do bloggers blog? Platform transitions within the historical Dutch blogosphere’, *First Monday*, 17(2)

Wilson, Chris (2014) ‘The Selfiest Cities in the World: TIME’s Definitive Ranking’, *TIME Magazine*, 10 March.

Zimmer, Michael (2015) ‘The Twitter Archive at the Library of Congress: Challenges for information practice and information policy’, *First Monday*, 20(7).

Richard Rogers is Professor of New Media and Digital Culture, Media Studies, University of Amsterdam. He is also director of the Digital Methods Initiative as well as the Netherlands Research School for Media Studies (RMeS).