Critical Web Archive Research

Anat Ben-David

Abstract Following the familiar distinction between software and hardware, this chapter argues that web archives deserve to be treated as a third category—memoryware: specific forms of preservation techniques which involve both software and hardware, but also crawlers, bots, curators, and users. While historically the term memoryware refers to the art of cementing together bits and pieces of sentimental objects to commemorate loved ones, understanding web archives as a complex socio-technical memoryware moves beyond their perception as bits and pieces of the live Web. Instead, understanding web archives as memoryware hints at the premise of the web's exceptionalism in media and communication history and calls for revisiting some of the concepts and best practices in web archiving and web archive research that have consolidated over the years. The chapter, therefore, presents new challenges for web archive research by turning a critical eye on web archiving itself and on the specific types of histories that are constructed with web archives.

1 Introduction

The field of web archiving and web archive research is maturing. A decade ago, most scholarly publications were concerned with questions characterizing the emergence of a new field of research and practice: How to archive the Web? Who is it for? In which ways does web archiving differ from archiving other digital or analog media? (cf. Costa and Silva 2010; Dougherty et al. 2010; Gomes et al. 2011; Niu 2012). Today, there is already a considerable amount of empirical research that no longer asks what web archiving is but instead uses the archived Web for answering various research questions, using a diverse set of methods. In recent years, the number of funded projects, publications, books, and conferences has grown from a handful to

A. Ben-David (✉)
Department of Sociology, Political Science and Communication, The Open University of Israel, Ra'anana, Israel
e-mail: anatbd@openu.ac.il

D. Gomes et al. (eds.), *The Past Web*,
https://doi.org/10.1007/978-3-030-63291-5_14

dozens. The last 2 years, in particular, have seen the publication of monographs and edited volumes on the topic and the establishment of *Internet Histories*, an international journal dedicated to studying the web's past (Brügger 2018; Brügger et al. 2018; Brügger and Laursen 2019; Brügger and Milligan 2018; Goggin and McLelland 2017).

From the theoretical perspective of the social construction of technology, to say that the field of web archiving and web archive research has matured is to point at technological closure and at a growing consensus shared by practitioners and researchers alike (Hård 1994). The professionalization of web archiving is evident in international collaborations and the development of standards. The establishment of international organizations such as the Internet Memory Foundation and the International Internet Preservation Consortium has contributed to the development of standards and best practices (Costa et al. 2017). Heritrix has become the default crawler used by most web archiving institutions (Mohr et al. 2004), the Wayback Machine and "Open Wayback" have become the default devices for replaying archived websites (Maemura et al. 2018), and the WARC file—which just celebrated its tenth birthday—is the standard file format (ISO 2009). In a similar way, there is also a shared awareness among web archivists that the breadth and depth of archival coverage of the Internet Archive differ from that of national web archives (Masanès 2006), that there are temporal inconsistencies in web archiving (Aubry 2010), and that current web archives do not handle duplicates well (Gomes et al. 2006). This consensus is shared by the web archiving research community, which has spent the past years sharing questions, issues, and methods (Schafer et al. 2016; Winters 2017). Recent work on tool development and standardization of methods is also a result of the important collaboration between web archiving institutions and researchers. Research-driven web services such as the Memento API (van de Sompel et al. 2010) and the Archived Unleashed toolkit (Milligan et al. 2019) are indications of this.

Despite the benefits of the ability to share standards, best practices, and knowledge across different communities, the maturation of web archiving as a research field, along with the technological closure of web archiving techniques, might also result in "black-boxing" of some of its processes. Since the fundamental questions have already been considered, researchers do not need to rethink the meaning and methods of web archiving every time they engage in a new research project. The problem with black boxing is that these processes gradually become taken for granted (Brügger and Milligan 2019). After the long process resulting in the consolidation of standards, best practices, shared methods, tools, and knowledge about web archiving and web archive research, there is also room for thinking critically about web archives and for rethinking some of their premises.

The purpose of this chapter is therefore to call for a more critical engagement with web archives. Thinking critically about the archived Web does not entail engaging in a righteous debate discerning right from wrong or discussing what ought to be better. Instead, I propose engaging in an epistemic debate and highlight some of the overlooked aspects of web archiving. Instead of asking "What are the best ways to archive the Web?" or "Why are web archives not widely used?", researchers could begin asking questions about the types of knowledge that web archives produce and reproduce, their embedded values and ideologies, their limits, and artifacts and politics.

To make the case for the necessity of critical web archive research, the following sections of this chapter are structured around three case studies that I conducted between 2014 and 2019. Each case study is based on one critical question about web archives and web archive research and is situated in a different geopolitical and temporal context.

2 Is the Wayback Machine a Black Box? Lessons from North Korea

When researchers view a snapshot of an archived website, they consider this snapshot as evidence of the web's past. The snapshot is an indication that this URL—along with its source code, content, and other elements—was part of the live Web at the time of archiving. But do we know enough about the circumstances that led to the archiving of this URL?

One way of addressing this question is to argue that it does not matter, as long as there is an exact time stamp attached to the archived snapshot. Another way is to acknowledge that the circumstances that lead to the archiving of a specific URL (and not another) are important in understanding how web archives might shape historiographical narratives and knowledge. For example, why are certain websites archived more frequently than others? Why, to date, does the Internet Archive have 60,821 snapshots of the website of the White House, compared to 2619 of the website of the Élysée? Who decided to increase the frequency of the archiving of the Olympic games website in the summer of 2016? Why did the frequency of the archiving of Egyptian newspapers not increase during the Arab Spring in 2011?

Asking these questions leads us to understand that, as in any other institutional archive, web archives may be biased and contain significant knowledge gaps. Therefore, one way of studying web archives critically is to try to find answers which account for inconsistencies. North Korea is a case in point.

North Korea might not be the first place that comes to mind when thinking about web archiving. Very little is known about the Internet in this secluded country. The North Korean Web is one of the smallest national webs: in 2016, a DNS leak in one of the country's root servers exposed that there were only 28 websites registered in the .kp domain. Although the DNS leak was treated as "breaking news", the scope of the .kp domain could have already been estimated using the Internet Archive, which had snapshots of most North Korean websites archived from as early as 2010. How did the Internet Archive "know" about the North Korean Web years before the leak?

Researchers make various assumptions about web archives. I, for example, assumed that one of Wayback Machine's crawlers must have captured the North Korean websites incidentally, by following links from other websites. But when we analyzed the sources that contributed snapshots, we found that knowledge about North Korean websites was mostly contributed to the Internet Archive by experts, archivists, and activists, rather than by automation (Ben-David and Amram 2018).

Another epistemic assumption about the Internet Archive is that web archiving is agnostic to geolocation and geopolitics, but through studying North Korean websites in the Internet Archive, we found that this is not the case. While the process of URL contribution is distributed (anyone can save a page to the archive from anywhere in the world), the archiving itself is centralized and based in the United States. Since there is partial access to North Korean websites from the United States, some of those websites could not be archived, even though they were on the Internet Archive's seed list. Put differently, the archivability of websites depends on geopolitics. The effect of geopolitics on web archiving leads us to the second set of critical questions that can be asked about web archives.

3 What Does the Web Remember of Its Deleted Past? Lessons from Yugoslavia

Yugoslavia is a country which was part of the web's history, but no longer exists. In 1989, the Socialist Federal Republic of Yugoslavia (SFRY) joined the Internet after the delegation of its country code Top Level Domain (ccTLD): .yu. Two years later, the country dissolved, and gradually, the countries that were formerly part of the SFRY received their own, new, national domains: Croatia and Slovenia were the first, and North Macedonia was the last. Throughout this time, the .yu domain continued to work—first as the official domain of the FRY and, then, as a historical digital remnant of both the Web and Yugoslavia's part of it (Ben-David 2016).

All these years of war, bloodshed and displacement are a crucial part of human history. These Yugoslav websites also documented a crucial part of the web's history, as it was considered "the first Internet War", involving online reporting and the spread of information warfare (Keenan 2001). But all of the digital remains of this important period are gone, due to unrelated Internet governance policies. In 2010, the .yu domain was removed from the Internet's domain name servers. This means that even if a .yu website is still hosted on a server, it is no longer part of the Internet root, and therefore cannot be found.

Thus, through the lens of the history of the .yu domain, the critical question to be asked about web archiving is: "What does the Web remember of its deleted past?" Of course, the Wayback Machine captured many of the .yu websites in real time. The problem was (and to some extent still is) that user access to web archives assumes that one knows, and subsequently types, a URL, to view its archived snapshots. Four years after the deletion of the .yu domain, it was nearly impossible to use the live Web to find Yugoslav websites. Subsequently, the Yugoslav websites that were archived in real time could not be reached, for all information about their past URLs was removed from the Internet.

Eventually, a list of URLs was found offline, which opened up a gateway to reconstructing a considerable portion of the archived Yugoslav websites in the Internet Archive. Yet, when I visualized the development of the hyperlinked structure of the reconstructed domain over time, I noticed that the domain became significantly interlinked only after the fall of the Milosevic regime, and most significantly after it became the domain of Serbia and Montenegro. That is, the structural evolution of the national domain indicates that sovereignty is inscribed into the politics of Internet governance and subsequently also affects the ability of the Web to remember its past. While the question of sovereignty is less significant for stable, wealthy countries, it seems that national web histories of countries in transition are particularly vulnerable.

The consequences of the inscription of sovereignty in web archives are even more grave for Kosovo, a country that, due to a Russian veto at the UN, does not have a ccTLD (Ben-David 2019a, b). That is, if it was at least possible to develop methods for reconstructing the Yugoslav Web from the Internet Archive through the domain suffix, it is nearly impossible to identify a Kosovar website on the live Web, and that has severe consequences for the preservation of Kosovar web history.

4 What Informs Web Archiving Policies? Lessons from Gaza

The premise of web archiving is that it captures discrete URLs in real time. Since preserving the entire Web is technically impossible and web archiving, in general, is costly, over the years most national web archiving institutions have developed policies that translate their mission to archive the Web into specific technical parameters. These policies often address issues such as the scope of archiving (full domain or special collections?), the boundaries of archiving (everything in the country code top-level domain? websites in a national language hosted elsewhere on the Web?), the frequency of archiving, and so on.

In most cases, this “translation” results in forming “seed lists”, or starting points from which web archiving begins. Given the technical complexity of web archiving at scale, it is almost impossible to change these seed lists in real time.

The results, however, are web archives comprised of distinct units: URLs that have been preserved as a result of a given policy and a specific method, at a specific point in time, and for a specific purpose, in order to preserve something that is relevant to a particular country. Are current methods for informing us about web archives sufficient if we are to use the archived Web as a significant source for historical research? What methods can be used to understand the impact of web archiving policies on shaping historiographical narratives, or to critique them?

Recently, my colleagues and I developed a method for building retrospective special collections of URLs around a past issue, or event, across various web platforms and national cultures (Ben-David 2019a, b). Apart from the technical challenges related to archiving the Web in retrospect, which are addressed elsewhere, the method aims to challenge the traditional sources that inform web archives. Most national web archives use seed lists as starting points for domain harvests. However, seed lists are agnostic to the wider context in which URLs are shared and discussed on the Web. With the growing platformization of the Web, information no longer travels across URLs but is rather confined to platform boundaries. Fixed seed lists are also agnostic to dynamic events that may coincide with routine crawls, but were not purposefully captured in real time. A cross-platform approach to web archiving addresses these problems, by incorporating the cultural context that comes along with how websites were distributed across various web platforms and at the same time taking into account cultural differences in URL sharing preferences. The case study we used for building a cross-platform archive was the 2014 war in Gaza.

The war, which lasted 50 days in the summer of 2014 and cost the lives of many Palestinians and Israelis, did not only take place on the ground. On social media, the fighting parties were heavily engaged in information warfare campaigns, and millions of users from around the world were involved in angry debates. News websites were reporting the events, pushing breaking news alerts around the clock, and on Wikipedia, edit wars were taking place about how to properly name and document the unfolding event. However, the majority of online activity relating to the war was not archived.

To reconstruct a cross-platform archive of the war, we used Wikipedia as the authoritative source for identifying the names of the war in 49 languages and used these as keywords for querying and scraping data and URLs from Twitter, YouTube, and Google Search. Using this method, we collected 118,508 unique URIs and relevant metadata, carbon-dated to the period of the military operation, in 49 languages from 5692 domain suffixes. Interestingly, we found significant cultural differences in URL sharing practices across platforms: while there are relatively few references in Arabic on Wikipedia and YouTube, Arabic language speakers mostly took to Twitter to discuss the issue and report the events. By contrast, URLs in Hebrew are mostly published by media outlets, which explains the relatively high proportion of these on Google and YouTube. We also found that some platforms are more prone to link rot than others—especially because of the role URL shortening services play in facilitating link sharing on social media.

These cultural and platform differences are crucial for informing us and thinking about web archives. Current web crawling methods are blind to the rich cultural and temporal dynamics that characterize the Web and are poor in contextual metadata. It would be useful for web archiving institutions to first identify and understand cultural and platform differences, before deciding on how, when, or where to archive the Web. A cross-cultural and cross-platform approach to web archiving also requires web archives to explore beyond their comfort zones. As we have seen with Yugoslavia, North Korea, Kosovo, and Gaza, the standard practice of thinking about web archiving from a national perspective might be a curatorial and institutional solution that stands in stark contrast to the global and networked structure of the open Web.

5 Conclusions

Web archives are the web's memory organs, and as such, they are breathing, dynamic, and constantly evolving. Consequently, web archives entail both a promise and a challenge for historical research. In this chapter, I attempted to take the promise with a pinch of salt, by arguing for the necessity of asking critical questions about web archives as epistemic agents: How is knowledge produced and by whom? What was not or could not have been preserved? What are the sources that inform web history, and how may each source shape a different historiographical narrative? To make the case for these critical questions, I presented examples from contested areas. Arguably, these sites of contestation invite us to think about web archiving critically, for it is at the periphery rather than the center where some of the assumptions we make when archiving the Web and when studying web archives no longer hold.

Critical web archive research may be useful to both researchers and practitioners of web archives: it may encourage them to think more reflexively about web archives as active agents, which have embedded values, biases, and politics, and about how web archiving techniques and policies are canonizing very specific ways of knowing the web's past.

References

Aubry S (2010) Introducing web archives as a new library service: the experience of the National Library of France. Liber Quarterly, Open Access journal of the Association of European Research Libraries, http-persistent

Ben-David A (2016) What does the web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain. New Media Soc 18(7):1103–1119

Ben-David A (2019a) National web histories at the fringe of the web: Palestine, Kosovo, and the quest for online self-determination. In: Brügger N, Laursen D (eds) The historical web and digital humanities: the case of national web domains. Routledge, Abingdon, pp 89–109

Ben-David A (2019b) 2014 not found: a cross-platform approach to retrospective web archiving. Internet Hist. https://doi.org/10.1080/24701475.2019.1654290

Ben-David A, Amram A (2018) The internet archive and the socio-technical construction of historical facts. Internet Hist 2(1–2):179–201

Brügger N (2018) The archived web: doing history in the digital age. MIT Press, Cambridge

Brügger N, Goggin G, Milligan I, Schafer V (2018) Internet histories. Routledge

Brügger N, Laursen D (2019) The historical web and digital humanities: the case of national web domains. Routledge

Brügger N, Milligan I (2018) The SAGE handbook of web history. SAGE

Brügger N, Milligan I (2019) Internet histories and computational methods: a "round-doc" discussion. Internet Hist:1–21

Costa M, Gomes D, Silva MJ (2017) The evolution of web archiving. Int J Digit Libr 18(3):191–205

Costa M, Silva MJ (2010) Understanding the information needs of web archive users. In: Proceedings of the 10th international web archiving workshop, 9(16): 6

Dougherty M, Meyer ET, Madsen CM, Van den Heuvel C, Thomas A, Wyatt S (2010) Researcher engagement with web archives: state of the art. Joint Information Systems Committee Report

Goggin G, McLelland M (2017) The Routledge companion to global internet histories. Taylor & Francis

Gomes D, Miranda J, Costa M (2011) A survey on web archiving initiatives. In: International conference on theory and practice of digital libraries. Springer, Berlin, pp 408–420

Gomes D, Santos AL, Silva MJ (2006) Managing duplicates in a web archive. In: Proceedings of the 2006 ACM symposium on applied computing, pp 818–825. ACM

Hård M (1994) Technology as practice: local and global closure processes in diesel-engine design. Soc Stud Sci 24(3):549–585

International Organization for Standardization (ISO) (2009) 28500: 2009 Information and documentation-WARC file format. International Organization for Standardization

Keenan T (2001) Looking like flames and falling like stars: Kosovo, 'the first internet war'. Soc Ident 7(4):539–550

Maemura E, Worby N, Milligan I, Becker C (2018) If these crawls could talk: studying and documenting web archives provenance. J Assoc Inf Sci Technol 69(10):1223–1233

Masanès J (2006) Web archiving: issues and methods. In: Web archiving. Springer, Berlin, pp 1–53

Milligan I, Casemajor N, Fritz S, Lin J, Ruest N, Weber MS, Worby N (2019) Building community and tools for analyzing web archives through datathons

Mohr G, Stack M, Rnitovic I, Avery D, Kimpton M (2004) Introduction to Heritrix. In: 4th international web archiving workshop

Niu J (2012) An overview of web archiving. D-Lib Mag 18(3/4)

Schafer V, Musiani F, Borelli M (2016) Negotiating the web of the past. Fr Media Res

van de Sompel H, Nelson M, Sanderson R, Balakireva L, Ainsworth S, Shankar H (2010) Memento: time travel for the web. Computer science presentations. Retrieved from https://digitalcommons.odu.edu/computerscience_presentations/18

Winters J (2017) Breaking in to the mainstream: demonstrating the value of internet (and web) histories. Internet Hist 1(1–2):173–179