Web Archives as Memoryware:

Critical Reflection on Sources and Methods for Web History

2019 International Internet Preservation Consortium General Assembly and Web Archiving Conference, Zagreb, June 7

Anat Ben-David

Good morning. I am very honored to be here today, and I am grateful to the organizing committee for inviting me to share with you some of my thoughts about web archiving and web archive research.

I began studying web archives in 2012, as a post-doctoral researcher at the University of Amsterdam. I was a member of the WebART project, whose goals were to develop retrieval tools for facilitating the use of web archives for scholarly research. When I started working on the project, my supervisor, Prof. Richard Rogers, half-jokingly warned me: "You are aware that you are entering a very small field." he said, "you can count the people who study web archives on one hand". He was right.

This conference's theme: "Web Archiving Community: Maturing Practice Together" is a reminder that the field has grown significantly ever since.

Over the past years, the number of funded projects, publications, books, and conferences has grown from a handful to dozens. This slide displays only a few examples of books that came out in the past two years–thanks to the vigor and dedication of Niels Brügger, who steadily pushed to have web archive research and Internet histories recognized as a scholarly field within communication and media studies.

If, seven years ago, most scholarly publications were concerned with how to archive the web, who is it for, and why web archiving differs from archiving other digital or analogue media – 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.

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 to technological closure, and at a growing consensus shared by practitioners and researchers alike.

The professionalization of Web archiving is evident in international collaborations and development of standards. The establishment of international organizations such as the Internet Memory Foundation and the IIPC, has contributed to the development of standards and best practices. Heritrix has become the default crawler used by most web-archiving institutions, the Wayback Machine has become the default device for replaying archived websites, and the WARC file -- which just celebrated its 10th birthday -- is the standard file format.

In a similar way, there is also consensus amongst web archive researchers: Many of us are aware of the differences in the breadth and depth of archival coverage of the Internet Archive compared to national web archives; Most of us are aware that both capture and replay of archived websites suffer from temporal inconsistencies; We know we have to deal with duplicates; and so on and so forth. This consensus is shared by the web archiving research community, who has spent the past years in sharing questions, issues, and methods. Recent work on tool development and standardization of methods is also a result of the important collaboration between web archiving institutions and researchers. Important web services such as the Memento API and the Archived Unleashed toolkit are indications of that.

Despite the benefits of being able to share standards, best practices and knowledge across different communities, the maturation of web archiving might result in black-boxing of some of its processes. Since most questions have already been considered, we do not need to rethink the meaning and methods of web archiving every time we engage in a new research project.

We do not have to worry about crawler settings because these just work so well, and the Wayback Machine; well, we really can't do without this marvelous invention, can we?

The problem with black boxing is that these processes gradually become taken for granted.

Now that we have standards, best practices, shared methods, tools and knowledge about web archiving and web archive research, perhaps the time has come to pause and rethink some of their premises.

Instead of asking: “what are the best ways to archive the web?”, or “why are web archives not widely used”? We can begin asking questions about the types of knowledge that web archives produce and reproduce; about their embedded values, ideologies; their limits; artefacts and politics.

The purpose of this talk 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, highlighting some of the overlooked aspects of web archiving.

To do so, I propose using the concept of memoryware as an analytical prism for the critical study of web archives. This analogy draws from the familiar distinction between hardware and software as objects characterizing computational media. While hardware relates to the material equipment used to construct electronic media, software relates to the programs and code used to operate them.

To talk about web archives as epistemic memory objects, I propose the term ‘memoryware’ as a third distinction. I did not invent this term, however.

While there is no official dictionary definition of the term, it is known as a synonym for broken tile mosaics, the art of piecing together tile shards and glazed chinaware. In African-American folk history, it also refers to specific memory practices, the art of using sentimental objects to create a unique, personal tribute to loved ones, used to decorate gravestones.

Thus, in the context of web archiving and web history, I propose using the term ‘memoryware’ to refer to the medium specificity of web archives as both the web’s memory organs, as well as to the specific historiographical practices that can be done while using web archives as primary sources. The analytical framework of understanding web archives as memoryware points to the double meaning of the term. On one hand, we can treat archived websites as sources for folklore. If memory jugs were created by cementing broken tile mosaics, buttons, and shards of other objects – we could view the archived website, or web archives as analytical units; as an amalgam of bits and pieces of sharded websites.

On the other hand, we can refer to web archives as memoryware, in the sense of the term’s analogy to other computational objects, such as hardware and software. For analytical purposes, I define web archives as memoryware to refer to the complex, hybrid and specific forms of preservation techniques, involving both software and hardware, but also crawlers, algorithms, policies, curators and users – through which the web’s history is both documented and constructed. I would argue that understanding web archives as memoryware in this sense of the word allows taking web archive research to new analytical heights. It allows us to ask questions about the technological imperative, as well as about the interpretive, ideological, analytical, and methodological implications of the use of this amalgam of bits and pieces, as our most trusted primary source for web historical research.

For the remainder of this talk, I invite you to join me on a geographical and temporal journey, below and off the grid. Each stop in this short journey will ask one critical question about web archives. When we reach our destination, I hope we will also have initial answers to these questions.Our first destination is North Korea, so please fasten your seatbelts.

***

What does North Korea have to do with Web archiving, anyway?

As far as we know, not much. Very little is known about the Internet in North Korea. 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 the fact that there were only 28 Websites registered in the .kp domain. Yet an examination of the archived snapshots of the North Korean Websites at the Wayback Machine revealed that all of them were already archived before the DNS leak – some of them, back in 2010. How did the Internet Archive 'know' about the North Korean Web years before the leak? This mystery leads us to our first critical question:

Is the Wayback Machine a black box?

That is, if we are to use archived snapshots as evidence, can we trace the specific reasons and circumstances that led to the archiving of these snapshots?

The answer to this question unravels fascinating, yet complex knowledge production mechanisms behind what we eventually perceive as archived snapshots (and hence, as evidence). There may be two identical versions of an archived website, but no two snapshots are alike, since the circumstances that led to their archiving reflect rich and multifaceted epistemologies, which may involve different actors, motivations, politics and interests. When we use archived websites for historical research, do we ever wonder how they got there in the first place?

Consider this quote from David Karpf, who describes the qualities of the Wayback Machine:

> We can think of the Wayback Machine as a "lobster trap" of sorts. Lobster trapssit passively in the ocean, placed in areas of strategic interest. From time to time, one can check the traps and see if anything interesting has come up. The Internet is similarly awash in data that may be of interest to researchers. We often want to make across-time comparisons. But without lobster traps, we are bound to go hungry, so to speak (2012, p. 648-649).

Karpf's "lobster trap" metaphor harbours assumptions about the Wayback Machine as a passive, rather than an active epistemic agent. The Web archive is passively "placed" in areas that others determine as strategic; and it is the role of others to determine whether or not it has "caught" something of interest. That is, web archive researchers make epistemic assumptions about the archived web that reflect a certain amount of trust in its technological architectures.

This level of trust is evident in the treatment of archived snapshots as facts. Despite recent contestation of web data as authentic sources, two of the web's non-commercial knowledge devices remain thus far relatively uncontested: Wikipedia and the Internet Archive.

Setting aside critique about certain bias in the editing of controversial entries and hidden power relations in the content management structure of the crowd-sourced Wikipedia, the fact that the history of all entries is both documented and transparent turns it into both an encyclopedia, and an archive.

But compared to Wikipedia, the Internet Archive's specific content-creation and distribution processes are less transparent. It would be easy to fall for the epistemic assumption of the 'lobster trap', and think that the Wayback Machine had captured the North Korean websites by chance. I, too, assumed that one of the Wayback Machine's crawlers captured them by following links from other websites. But the more we looked into the matter, the more we realized that the process of knowledge production performed by automation is far less significant than human knowledge and intervention.

To figure out how the Wayback Machine came to 'know' about the North Korean websites and to archive them, my colleague Adam Amram and I used the 'provenance' feature that was added to the Wayback Machine in 2016. This feature provides information about the 'organization' that contributed the snapshot, as well as the 'collection' in which that snapshot is found. We scraped all provenance information for every archived snapshot of every North Korean Website, and mapped the results.

We found that, next to the Internet Archive's crawlers, proactive human contribution plays a significant role. For example, Mark Graham, director of the Wayback Machine, is the 'organization' that contributed the most snapshots. Other 'organizations' are the national library of Australia, two curators working at the Internet Archive, and the Archive Team - a collective founded by Jason Scott in 2009, comprised of programmers, archivists, writers, and activists dedicated to preserving digital history. So at least quantitatively, the contribution of North Korean websites to the Internet Archive by human experts, trained archivists and activists is far greater than the contribution of automated crawls based on initial seed lists.

Another epistemic assumption about the Internet Archive, is that the archive's location does not effect archival coverage. But examination of North Korean Websites on the Internet Archive revealed 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 this is where geopolitics come into the picture: the archivability of websites may depend on diplomatic ties, and internet censorship policies in different countries.

Accessibility rates to North Korean websites vary across countries, depending on their political ties with Pyongyang. Compared to a 100% accessibility rate in Russia, only 50% of North Korean websites can be accessed from the US. Geolocation, then, also effects the production of historical facts.

***

If there are web archivists from national libraries in the audience, at this point, you must feel relieved – you might be thinking, "oh, this might apply to the Internet Archive but is no concern of ours. Since we only preserve websites in our national domain, we do not have to be concerned about diplomatic ties."

But since we began talking about the geopolitics of web archiving, let's continue in our imaginary geographic and temporal tour, and take a flight from the Wayback Machine's servers in San Francisco over here, to Croatia.

Now let's time travel back to 1989. Croatia was still part of the Socialist Federal Republic of Yugoslavia, and it had just received a fresh Country Code Top Level Domain: .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, North Macedonia was the last. Throughout this time, the .yu domain continued to work – first as the official domain of FRY, and then, as a historical digital remnant of both the Web and Yugoslavia's part.

These years of war, bloodshed, and displacement, are a crucial part of human history. The .yu 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.

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 System. 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.

The history of the .yu domain, brings me to my next critical question, which is:

<u>What does the web remember of its deleted past?</u>

The answer to this question sheds light on two epistemic processes that relate to the geopolitics and temporality of web archives: The first is that web archiving inherently depends on the politics of the Internet's DNS system. And the second is that the archived web is temporally dependent on the live web.

We can also think of the DNS as memoryware of sorts, since it translates IP numbers into mnemonic addresses. As a hierarchical and universal system for the resolution of Web addresses, the DNS is the Internet's most strict authenticator of sources: HTTP requests of Web addresses incompatible with the DNS will not resolve. At the same time, the DNS is also the Internet's most strict authenticator of nation-states. The DNS is managed by ICANN, which delegates ccTLDs to countries enlisted in ISO-3166-1, the list of the official names of countries and territories recognized by the UN, and their two-letter suffix. As new countries are added to the list, their newly delegated ccTLDs are added to the DNS and subsequently emerge on the live Web. But when countries dissolve, a removal of a ccTLD from the DNS consequently deletes the possibility of resolving its historical addresses on the live Web. The other side of the protocol of mnemonics is thus permanent memory loss.

Arguably, the dependence of the live Web on the DNS, and consequently of the Web archive on the live Web, inscribes sovereignty and stability into Web archives and national Web history. Sovereign countries whose historical ccTLDs have expanded over the years enjoy the benefit of the enduring proximity between the live Web and its archiving. At the same time, such inscription of sovereignty jeopardizes the Web histories of unstable domains or non-sovereign states and peoples, whose digital pasts are dotted with rupture and deletion.

Of course, the Wayback Machine has 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 snapshot. Four years after the deletion of the .yu domain, it was nearly impossible to use the live web to find Yugoslavian Websites.

While I do not have enough time to go deeply into the process of reconstruction, I managed to find an initial list of .yu websites that were captured by a Serbian Wikipedian, about a month before the domain was removed, and used this list to reconstruct a considerable portion of the .yu domain from the Wayback Machine. But if we only look at the visualization of the domain over time, we already can see the inscription of sovereignty into the archived Web. 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.

The consequences of the inscription of sovereignty in Web archives are even more grave. Due to a Russian veto at the UN, Kosovo does not have a country code top level domain. Therefore, if it was at least possible to develop methods for reconstructing the Yugsolav Web from the Internet Archive through the domain suffix, it is nearly impossible to identify a Kosovar website in the live web, and that has severe consequences on the preservation of Kosovar web history.

***

So far, we've travelled in contested countries and areas. The final stop on our journey will not be different. We are jump-cutting to Gaza in the summer of 2014.

The critical question I would like to ask here, is:

<u>What informs web archiving policies?</u>

In the summer of the 2014, Israel and Hamas were engaged in a violent conflict officially dubbed by the Israeli military "Operation Protective Edge", but widely acknowledged as the

2014 War in Gaza. During the 50 days of the operation, the IDF carried out more than 6000 air strikes, alongside a massive ground operation, which took place between July 17 and August 5. About 70% of Gaza's population evacuated their homes during the attacks; hundreds were killed and thousands injured. Hamas, for its part, launched thousands of rocket attacks on Israeli cities, causing injuries, casualties, and distress among Israeli civilians.

Imagine the Web that year during the war: Millions of users from around the world were debating it on social media. The Israeli military and Hamas also used social media to engage in information warfare. News website were reporting the events, pushing breaking news alerts around the clock, and on Wikipedia, editing wars were taking place on how to properly name and document the unfolding event.

These abundant and dynamic communication traces provide important documentation of a significant event, both for the Palestinians and Israelis, and internationally.

But they are not part of the archived web.

During that time, web archiving crawlers at the Internet Archive and at national libraries around the world were performing their routine harvests of websites, based on their regular settings. Unintentionally, they may have captured some of the URLs that were published during the war, if these fell under the crawler settings, or if users actively saved them to the Wayback Machine. But the majority of the online activity related to the war – especially that which took place on social media platforms, outside the realm of web archiving crawlers – was not archived.

We are all aware of the web's ageing problem, also dubbed as 'link rot', or 'web decay'. If web materials are not preserved in real time, they will most likely vanish within two to four years.

Over the past decade, the web has also undergone platformization, meaning that the majority of content no longer travels across a decentralized network of hyperlinks, but is confined to the walls of dominant social media platforms such as Facebook, Twitter, and YouTube, and is served through centralized data services.

Archives, museums and galleries are filled with objects which may have existed for decades before being curated, archived or preserved. Unlike these objects, the preservation of the web has to take place in real time.

Borrowing its terminology from photography, the verb used to describe web archiving is 'capturing', and the metaphor to describe an archived web object is a 'snapshot', or a 'memento'. These metaphors illustrate the assumption that since the web is a dynamic medium, its archiving locks in a specific moment in time, one which later serves as evidence that the archived website – which may have changed dramatically afterwards—was part of the live web at the moment of capture.

Four years after the War on Gaza, can its web presence still be preserved?

My colleagues and I have attempted to develop a method for building retrospective special collections of URLs around a past issue, or event, across various web platforms and national cultures.We used Wikipedia as the authoritative source for building the cross-platform and cross-national collection, and scraped the URLs from the Wikipedia pages describing the war in 49 languages. We also used the names of the Wikipedia pages in 49 languages as keywords, and subsequently queried and scraped further data and URLs from Twitter, YouTube, and Google Search.

We further used the CarbonDate Api developed by the Web Science and Digital Libraries Lab at Old Dominion University, to determine that the URLs we captured were indeed published during the war, and used the save page now API to batch push all of them to the Wayback Machine.

The results are a retrospective archive comprised of 118,508 unique URIs and relevant metadata, carbon-dated to the period of the military operation, in 49 languages and 5692 domain suffixes. The Israeli domain harvest, which was performed for the National Library of Israel by the Internet Archive, contained less than 1% of the URLs we collected in the retrospective collection.

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 their relative high proportion on Google and YouTube. We also found that some platforms are more prone to link rot than others – especially due to the role URL shortening services play in facilitating link sharing on Social Media.

I argue that these cultural and platform differences are crucial for informing and thinking about Web archives. To date, the majority of web archiving institutions use web crawlers as a technique to perform large-scale, real-time web archiving of the web (the Internet Archive), or of national webs (national libraries). Yet web crawling captures URLs indistinctively. It is a practice blind to the rich cultural and temporal dynamics that characterize the web, and is poor in contextual metadata. In most cases it is unable to cross the walled gardens of social media platforms. Milligan and Ruest suggest distinguishing between ‘Gate keeping’ and ‘bottom up’

approaches to web archiving curation. In line with this distinction, web archiving institutions must first attempt to understand these 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 that web archiving explore beyond 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. Put differently, if we remain too comfortable trusting our matured standards and practices, we may fail to notice that the ship has sailed.

Having said that, web archiving in retrospect is also proposed as a way to redress the situation: It can be used by archiving institutions to correct archival divides, or to build inexpensive special collections after the fact. It may also be used by citizens, researchers and activists to counter hegemonic narratives which are deliberately, or arbitrarily, constructed by existing web archives.

***

Why did I take you on this journey, and ask these questions about web archiving in contested areas? I did so because these are the places where some of the assumptions we make when archiving the web and when studying web archives, no longer hold. We can continue studying national webs and characterize linking practices, link rot, and evolution. But we can also ask different questions. Unsurprisingly, we find ourselves back at the point of departure.

Web archives are the Web's memory organs; and as such, they are breathing, dynamic, and constantly evolving. If we are to treat Web archives as memoryware, there are two paths we

can take: We can continue treating web archives as bits and pieces; as digital shards of the web's past, and continue to use the historiographical methods we are familiar with; trying to cement the pieces together. Or, we can take the second path and treat web archives as active agents, with embedded values, biases and politics. As web archiving institutions and practitioners, we can do the same: either continue to collect bits and pieces, and leave it for future researchers and users to decide what to do with them; or think more reflexively on how web archiving techniques and policies are canonizing very specific ways of knowing the web's past.

Thank you.