Ben-David, A. & Amram, A. (2018, Forthcoming). Computational Methods for Web History. In: Brügger, N. & Milligan, I. (eds.) The SAGE Handbook of Web History. London: Sage (Accepted version)

Computational Methods for Web History

Anat Ben-David and Adam Amram

Introduction

In light of the exponential growth in digital data characterizing the 21st century, future historians of our time will have to rely on born-digital materials as primary sources for establishing historical facts. Yet born-digital materials challenge historians' well-established source criticism techniques used for establishing facts based on the authenticity, authorship and authority of documents, for they are ephemeral, immaterial, fragile and easy to manipulate. For example, the content of websites can be easily modified, tweets are frequently deleted, the number of social media comments and likes can be artificially boosted through click farms, and dubious sources spreading misinformation can be disguised as reliable news organizations. With the commercialization of the web, more than ever before, web data is primarily proprietary, and therefore subjected to platforms' policies and constraints.

Compared with the fragility and contestation of born-digital data, web archives are one of the last non-commercial knowledge devices that can be used to establish historical facts from web data. Web archives capture snapshots of websites at a specific point in time and preserve them for eternity. However, despite the fact that archived websites are stable, reliable, public and non-commercial digital primary sources, they are difficult to study with digital and computational methods.

In recent years, a 'computational turn' is advocated as a paradigmatic shift in the Digital Humanities. Arguably, such justification seems unnecessary for the web, for there is no need for a 'turn' if the web is a computational medium to begin with. On its face, many of the computational techniques already in use by digital humanists and web researchers (such as methods for text and image analysis, geo-mapping, or network analyses) can be easily used with archived web materials. The question remains, if web archives are conceived, curated and distributed in digital form, why are they so difficult to study computationally?

One of the reasons accounting for the difficulty in using computational methods on the archived web is the distinctive ontological status of archived websites, compared with other digital materials. According to Niels Brügger, web archives differ from other digital material, whether they are analogue documents that have been digitized (for example, scans of historical newspapers, letters, or books), or born-digital documents (such as copies of contemporary

newspapers, letters, books and other documents that were originally produced in digital form). Termed by Brügger (2016) as 're-born digital materials', the archived web is a static representation of an ephemeral medium, which would not have existed prior to its archiving. Brügger outlines further challenges regarding the unique characteristics of the archived web, including, among others, the fact that differences in the technical settings and collection policies of web archiving institutions may result in different archived versions of the same websites. In addition, there are difficulties in representing the temporal aspects of the archived snapshots, and in deciding which archived version of a given website should be used, where there is either scarcity of versions or a multiplicity of nearly identical versions from similar time stamps. Put differently, web archives are not just another repository of digital documents, but call for particular methods that would take into account their unique characteristics.

In this chapter, we argue that the use of computational methods for web archive research is possible – and even desired – provided that the methods, tools and techniques are adapted to the specific characteristics of web archives, and the challenges involved in studying them. To support this argument, we begin with an overview of the theoretical justifications for the use of computational methods in Computational Social Sciences and Digital Humanities. Thereafter, we discuss specific methodological challenges put forward by web archives, and how computational methods may be helpful in overcoming some of them. We further outline four computational techniques – drawn from our previous research projects – that illustrate ways in which computational tools have enabled us to use the Internet Archive as a primary source, and to answer web historical questions. Finally, we discuss the limits of the computational methods approach.

Theoretical Background

Broadly defined, the term 'computational methods' relates to the application of analytical methods and techniques – originally developed by computer scientists – to answer research questions in other disciplines such as the humanities and the social sciences. It is fair to say that these mathematical techniques, involving data processing, numerical analysis, simulation and modeling, algorithms, visualization, artificial intelligence and other forms of computation, are agnostic to disciplinary boundaries, as long as the data is digital. For example, a computer program designed to identify topics in a large corpus of texts can be applied to analyze a corpus of poems, of historical manuscripts, or of corporate emails.

However, the introduction of computational methods in the humanities and social sciences has been far from trivial. David Berry (2011) describes a 'computational turn' in digital humanities as a result of three paradigmatic waves: in the first, computational techniques were

primarily a means to digitizing texts and to building infrastructures for digital archives, which, in turn, were analyzed by humanists as they would analyze any other non-digitized text; the second wave looked at the digital historian's toolkit as one that can be applied not only to digitized texts, but also to any other 'born-digital' data. Finally, the third wave turned its gaze to computation itself – the mediation of computer code and the digitality of the objects it works with – as producing new ontologies and effecting epistemic changes.

Indeed, in the past decade, advances in computational technologies opened new possibilities for storing and processing data at a scale which was unimaginable only a few years ago. The analytical possibilities that accompany new computational analyses promise new kinds of knowledge that were impossible before. Big data analysis and computational methods are termed the new historians' 'macroscope' (Graham et al., 2015), alluding to the revolutionary impact the invention of optical instruments such as the microscope and telescope had on the natural sciences in the 17th century. This promise is epitomized in two papers published in the journal *Science*. In the first, Michel et al. (2011) put forward the notion of 'culturomics', involving computational analyses of large volumes of literary texts to quantitatively investigate cultural trends. Resonating Moretti's distinction between close and distant readings (2007), the researchers computationally analyzed a corpus of digitized texts containing about 4% of all books ever printed from 1800 to 2000, and demonstrated that this approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. The second paper discusses the application of a computational approach in the social sciences (Lazer et al., 2009). Here too, the researchers argue that the increasing datafication of almost every realm of human activity paves the road to new research opportunities that can identify patterns, trends and proximities in huge and diverse datasets, and draw new insights in many fields in the social sciences.

Going back to Berry's definitions of three waves of digital humanities, it can be argued that the study of the web and its history can be situated between the second and the third – by bringing together the humanist's and the social scientist's research questions and toolkits, and by studying born-digital texts, as well as data about almost any aspect of human activity. On the face of it, arguments for adopting a computational approach for web history do not require further justification, as they draw from both the digital humanities and the computational social sciences. As Graham et al. note:

While big data is often explicitly framed as a problem of the future, it has already presented fruitful opportunities for the past. The most obvious place where this is true is archived copies of the publicly accessible Internet. The advent of the World Wide Web in 1991 has had revolutionary effects on human communication and organization, and its archiving presents a tremendous body of non-commercialized public speech. There is a lot of it, however,

and large methodologies will be needed to explore it. It is this problem that we believe makes the adoption of digital methodologies for history especially important. (2015: 27)

The ability to perform new computational analyses at scale seems to be a unified justification for the adoption of computational and digital methods for history, social sciences and web research. Nevertheless, web historical research – with the archived Internet as its primary source – still posits somewhat different justifications for the use of computational methods, since the web is a computational medium to begin with. If the 'computational turn' in the digital humanities and the social sciences began with the digitization of printed texts on one hand, and with the 'datafication' of social phenomena on the other, for the web – there has never been a need for such a computational turn.

The question then remains: if there is such a structural fit between the web as a computational medium and computational methods for studying its history, why are computational methods not widely used in web historical research?

The answer to these questions may be found in Brügger's alternative periodization of Digital Humanities, which focuses on the evolution of digital material, rather than on research practices (2016). Compared with digitized and born-digital materials, Brügger argues that web archives should be considered re-born digital materials, which grants it a distinct ontological status and calls for a methodological treatment that takes its unique characteristics into account. Therefore, the apparent structural fit between the 'digitality' of the archived web and digital methods designed for studying born-digital material, such as the live web, is misleading.

In the next section, we follow Brügger's argument about the distinctiveness of the archived web as re-born digital material, by tracing the historical roots of the use of computational methods in web history, and by outlining specific challenges that hinder the adoption of computational methods in web archival research. We then offer possible computational solutions to some of these challenges.

Roots of Computational Methods in Web History

The establishment of the Internet Archive in 1996 marked the emergence of web archiving practices, now shared by many institutions, organizations and researchers around the world. Five years later, snapshots of archived website became available for browsing and viewing through an interface developed by the Internet Archive and aptly labelled the 'Wayback Machine'. In many ways, the introduction of the Wayback Machine has paved the road for both qualitative and quantitative analyses of the archived web.

Historically, however, computational and quantitative methods for studying web archives were the realm of information scientists, and not historians, media scholars or social

researchers. The early information science literature on web archives offered computational tools for improving retrieval of archived web content, for example by using metadata (Rauber and Bina, 1999; Rauber and Merkl, 1999). Researchers have also started publishing results from longitudinal studies analyzing the structure and persistence of web pages, based on archived web data (Koehler, 2002). After the turn of the millennium, studies in information retrieval created large temporal web collections for the development and evaluation of retrieval techniques for the temporal web (Baeza-Yates et al., 2004; Bordino et al., 2008; Chen and Roy, 2009; Chung et al., 2009). Historical analysis of the dynamic evolution of websites' markup style was picked up by information scientists much later, less as a means for historical research and more as a method for evaluating the retrieval effectiveness of the modern web (Gyllstrom et al., 2012).

Meanwhile, social science researchers – mainly political scientists, and later on web historians and other media scholars – started using web archives for studying social and political phenomena on the web. The early methods used for the social studies of archived web materials were either manual, or only semi-automated. Researchers in the Netherlands, for example, started archiving websites of Dutch political parties and electoral campaigns (Voerman, 1998). In a series of publications, Schneider and Foot put forward the method of web sphere analysis, as 'a framework for web studies that enables analysis of communicative actions and relations between web producers and users developmentally over time' (Schneider and Foot, 2005: 2). The method involves dynamically selecting and archiving a set of web pages around a theme or an event, web pages which are subsequently analyzed by triangulating hyperlink, content and qualitative analyses (Foot and Schneider, 2010; Foot et al., 2003; Schneider and Foot, 2004).

Gradually, along with the widespread establishment of web archiving initiatives at national libraries, as well as with developments in the ways the web was structured and organized (i.e. through search engines) and with methodological developments in the fields of Internet research (such as automated hyperlink analysis and web scraping), digital and computational methods for the study of web archives began to emerge (Rogers, 2015).

For social scientists and historians who engage with Internet histories, web archives enabled very specific types of research based on a single archived web page as unit of analysis. Rogers describes one such research scenario as website histories – performed through screencast documentaries. Rogers claims that such a research scenario is native to the medium, as it 'makes explicit what the Wayback Machine implies, with its invitation to tell the history of a website and through it the history of the web' (Rogers, 2017: 90).

Following the idea of repurposing the web's natively digital objects and organizing devices to conduct 'natively digital' web research, Rogers and researchers from the Digital Methods Initiative also developed tools that output a list of direct links to all archived snapshots

of a given URL[1], and which extract a historical hyperlink network from archived snapshots of a set of URLs[2]. These tools have been used by Weltevrede and Helmond (2012) to study the evolution of the Dutch blogosphere.

It is only around the time of the emergence of the 'computational turn' in digital humanities and computational social sciences that large-scale historical analyses of web archives became more widespread. This has been primarily facilitated by emerging collaborations between academic researchers, digital libraries and web archives, such as the Big UK Domain Data for the Arts and Humanities (BUDDAH) project (Winters, 2017), the Netherlands Web Archive Retrieval Tools (WebART) project (Huurdeman et al., 2013) and the Danish Probing of a National web project (Brügger, 2017). These projects – all initiated around 2013 – aim to highlight the value of web archives as a source for arts and humanities researchers, develop access tools and methods for research, and provide research scenarios – as well as research findings – of historical studies of a national web archive as a (huge) single unit of analysis of several terabytes, which is highly dependent on computational analyses as well as on infrastructures suitable for big data analysis. In line with these developments are recent calls to think of computational methods as the new toolkit of the web historian (Milligan, 2012, 2016). Despite this, it should be noted that, to date, the ability to access and process an entire national web archive as unit of analysis is reserved to a handful of researchers who collaborate with national libraries or with the Internet Archive, while other members of the research community need to resort to available resources such as the Wayback Machine, which, as previously noted, is designed for viewing the history of a single website, rather than for performing large-scale data analyses. The following section further elaborates the challenges that hinder the widespread application of computational methods to archived web materials.

Challenges in applying computational methods for web history

Though the application of computational methods to study web history has become more widespread in recent years, most of the published literature on the use of web archives for historical research focuses on technical and theoretical challenges limiting the application of computational and digital methods for web history (Brügger, 2013; Dougherty et al., 2010; Milligan, 2016). These challenges can be organized analytically around four topics: issues of access, the limits of current interfaces to web archives, and problems of contextualization and completeness of archived web materials. All four challenges affect the historian's ability to make sense of the primary materials. Below we briefly discuss these challenges that hinder the application of computational approaches, or 'distant reading', to web archives (Lin et al., 2014).

Access

User statistics of web archives hosted by national libraries show relatively low access rates. To name two examples, both the French and Danish web archives keep full domain harvests of the .fr and .dk domains, respectively, under legal deposit law, and access is restricted to researchers. In 2012, the National Library of France reported 30–50 monthly consultations of the web archive, whereas in Denmark only 20 researchers received access between 2007 and 2012 (Schostag & Fønss-Jørgensen, 2012; Stirling et al., 2012).

There is a growing divergence between the technological ability to preserve and archive the web, and the legal barriers related to copyright and privacy that hinder online access to archived web materials. With the exception of the Internet Archive and the national web archives of Japan and Portugal, most web archiving initiatives held by national libraries can only allow offline access, and only to those physically located in the libraries' reading rooms. As a consequence, only a handful of researchers who actively collaborate with national web archives or with the Internet Archive are able to conduct large-scale computational analyses of the archived web[3].

Interfaces

The problem of legal access to archived web materials is coupled with a problem of existing access tools and interfaces to web archives. The design of the Internet Archive's Wayback Machine – currently the dominant access interface to most institutional web archives around the world – reflects the perception that the unit of analysis is a single website, and that archives are best consulted, or browsed, one website at a time (Brügger, 2012a; Rogers, 2013, 2017). The single-site approach hinders researchers from increasing their analytical scope from one page to a collection of websites, or even the entire archive (Ben-David and Huurdeman, 2014; Milligan, 2016).

User studies indicate that most users (whether they are journalists, litigators, lay people or academic researchers) expect to consult web archives in the same fashion they consult the live web: through search (Ball, 2010; Costa and Silva, 2009, 2011; Jatowt et al., 2008; Meyer et al., 2011; Ras and van Bussel, 2007). Although several web archives are already developing full text search interfaces, some of which were designed specifically for data analysis purposes (Holzmann et al., 2017), these have still not been widely implemented.

Indeed, current web archiving infrastructures are not well-suited for computational research. The technical infrastructure of most web archives – the Internet Archive among them – was built before the 'computational turn'. Arguably, web archiving infrastructures are also

mirrors through which the history of the web can be studied. The Wayback Machine's slogan, 'surf the web as it was', denotes that it was conceived at a period in the web's history when 'browsing' and 'surfing' were the common mode of engaging with the web, before the takeover of the search paradigm (Ben-David and Huurdeman, 2014). From an infrastructural point of view, modern technologies designed to process petabytes of data, such as Hadoop Distributed File System (HDFS) and Hadoop MapReduce, are underutilized. The lack of those contemporary big data technologies as infrastructure for web archive preservation and access makes it difficult for researchers to perform 'distant reading' and macro analytics on web archives (Lin et al., 2014).

Contextualization

Another critique often made about the archived web as primary source for historical research is a problem of contextualization. As Featherstone (2006) argued, digital storage allows for expanding the boundaries of the archive, as well as the boundaries of what is considered worthy of archiving. Yet this poses tremendous difficulties for the traditional practices of source appraisal and provenance, which are central principles in archival sciences. Web archives may include millions of web pages, yet the archived snapshots lack significant contextual information about the wider media ecology in which the website operated when it was archived, and lack significant provenance information (Ben-David and Amram, 2018; Dougherty et al., 2010; Milligan, 2016; Rogers, 2013). While web archives keep the seed lists and crawl logs of the archiving process, this important metadata is usually not made available to researchers. As a result, little is known about the circumstances of archiving a specific website at a particular point in time, about the specific archiving method, or whether or not a specific website has been archived deliberately (as part of a seed list), or serendipitously (as a linked website to a seed list) (Huurdeman et al., 2015)[4]. Even though the lack of sufficient contextual information and metadata limits the ability to draw significant findings from large-scale analyses of archived websites, as we soon demonstrate, a 'distant reading' of web archives may be fruitful in re-introducing some of the missing context to archived web materials.

Completeness

Completeness is another issue web historians need to address when working with web archives as primary sources, and where computational methods may be of crucial relevance. Niels Brügger (2012b) poignantly reminds us that web archives are both incomplete – since many pages are not archived – and too complete – since there may be many duplicates and archived versions of the same page. Making sense of archived web materials is therefore aided by assessments of the degree of their completeness, or incompleteness. Here, completeness can be

measured as an evaluation of the archiving process itself (what has been preserved and what has been lost), or of the outcome of the archiving process (which sources are missing from a given corpus of archived websites, and whether or not the archiving of websites is complete). For example, a study by Thelwall and Vaughn (2004) shows significant differences in the archival coverage of national webs on the Internet Archive. Research by Ainsworth et al. (2011) shows that an estimated 35%–90% of websites have at least one archived copy; however, these figures are challenged by Hale et al. (2017), who found that the archival coverage of the popular website TripAdvisor is only 24%; Alkwai et al. (2015) estimated the completeness of archived snapshots of Arabic websites on the Internet Archive by comparing them with a list of URLs that were found in web directories. Interestingly, Huurdeman et al. (2015) put forward the notion of the 'aura' of a web archive – representations of unarchived content that are represented by hyperlinks and anchor text data from archived URLs of a national web archive.

While such assessments of archival coverage and completeness are crucial for contextualizing web archive materials for historical research, they can only be performed computationally, since it is necessary to assess the whole volume of a given archive (and, as is the case with the studies referenced above, estimate the size of the entire web) in order to indicate or infer what parts of it may be incomplete or missing.

To summarize this section, the relatively low usage rate of web archives for historical research, and the adoption of computational methods for this type of study, can be explained as a chicken and egg problem. On the one hand, there is growing scholarly interest in accessing and using archived web data for various types of analysis. While there are emerging initiatives to develop frameworks for allowing access, data extraction and analysis of web archives[5], limited access to most existing web archives, along with the limits of interfaces and processing infrastructures, and lack of sufficient contextualization information about appraisal, provenance and completeness, drive many historians away from using web archives for web history, and especially from applying various data analysis methods that would be native to the medium. In the following section of this chapter, we attempt to evoke the web historians' interest in applying simple computational methods when using web archives for historical research. This is not proposed as a solution to the existing challenges described above; rather, such simple computational methods can be seen as a means to allow for distant readings of archived web materials, by stretching the limits of existing access channels, interfaces and infrastructures of web archives.

The techniques described in the following section are based on simple scripts specifically designed for researchers who may not be well versed in computational methods, and who do not have institutional access to web archives as big data corpora. This is provided so as to assist

more researchers in attaining as many analytical benefits as possible, when using the Internet Archive's Wayback Machine as their entry point for conducting research on the web's pasts.

Computational solutions to the challenges of access, interface, contextualization, appraisal and completeness

In the previous sections of the chapter, we reviewed some of the literature that applied a computational approach to studying web archives from diverse disciplines and perspectives, ranging from assessments of the archives' coverage, to the characterization of national webs. In this section, we report on findings from our own work, by outlining specific and simple computational techniques that we developed and applied to web historical research. These techniques are designed to overcome specific challenges related to archived web materials and their current interfaces, as outlined above. Furthermore, they significantly increase the scope of analysis enabled by the Wayback Machine. While these methods are tailored to specific questions, they serve here as a proof of concept, which can then be adapted and used for other research purposes. Above all, they are listed here as a means to lower the threshold preventing web historians from engaging in critical, born-digital, web historical research, such that can critique the historical, technical and infrastructural history of archived websites, and that can further contextualize them as primary sources for historiography.

Our previous work has primarily focused on the Internet Archive and the Wayback Machine as a primary source. The research questions – related to the history of national webs – were not computational in nature; rather, we used computational methods where a methodological challenge hindered us from increasing the scope of the analysis or from gaining wider contextual knowledge about our corpus. Below, we outline four research scenarios where the use of relatively simple Python scripts has helped us enhance the utility and scope of historical research with the Wayback Machine.

1. Increasing the Scope: Finding more URLs

As previously mentioned, the Wayback Machine allows for viewing and browsing archived snapshots, but does not lend itself easily to analyses that wish to extend their scope beyond the single page as unit of analysis. Furthermore, in order to use the Wayback Machine, users are expected to know the URL they are looking for[6]. But what if one does not know which URL to look for? And what if it is impossible to know the URL address of historical websites that can no longer be found on the live web?

This is the case of the .yu domain, the historical country code top-level domain of the former Yugoslavia, which was entirely removed from the Internet's Domain Name System (DNS) in 2010. DNS is the hierarchical universal system responsible for the resolution of Web addresses from IP numbers. As such, if the IP addresses associated with a given domain are removed from the DNS, they cannot be resolved, even if the websites are still hosted on a server connected to the Internet.

Our research question was relatively broad: what does the web 'remember' of its deleted past? Is it possible to reconstruct from the Wayback Machine a deleted national web, if the live web cannot disclose any .yu URL as a starting point to the archive (Ben-David, 2016)? To answer this question, we used the hyperlinked structure of archived websites to our advantage. Using lists of seed URLs which were obtained from an expert, we used a python script that fetched all the seeds from the Wayback Machine, and snowballed their outlinks, searching for more .yu URLs[7]. We iterated the method several times until no new .yu URLs were found. This outlink extraction method enabled the scope of the analysis to be increased from a seed list of about 4,000 hosts, to 17,460 unique websites in the .yu domain, estimated to comprise 53% of the registered addresses in the .yu domain at its peak. The partial reconstruction does not necessarily indicate that the undiscovered portions of the domain represent a limitation of the method; instead, it may indicate that almost half of the domain was not archived by the Internet Archive, as we will soon explain.

With the extracted dataset, we performed historical network analysis (a method outlined in Chapter 10 of this *Handbook*; see also Weltevrede and Helmond (2012) and Hale et al. (2014)) to examine the strength of ties within the .yu domain, as well as among national domains of former-Yugoslavia republics. Furthermore, the data we extracted from the Internet Archive allowed us to ask further questions about the history of the .yu domain, as well as about the appraisal of the Internet Archive as a primary source for web history, as outlined below.

2. Source Critique: Comparing Web Cultures of Seed Lists

As noted by Rogers (2017), archived websites face scrutiny as a primary source for historiography, and for legal evidentiary purposes. Such scrutiny is realized through traditional appraisal of sources in historiography, such as determining the originality of a document, the authenticity of authorship, and the source's reputation. While web archives complicate all of the above (Brügger, 2012a; Rogers, 2017), the question is which techniques can be used to appraise archived websites as sources? While the screencast documentary of a single website is one possible venue for source criticism, we put forward a technique for appraising lists of URLs, as starting points or seeds for archival web research.

Historical analysis of archived web materials that extends beyond the single website as its unit of analysis greatly depends on a seed list of initial URLs from which the corpus is created. Different URL lists may lead to very different corpora, even though they may be seen as covering the same topic. For example, a study by Mataly (2013) compared corpora created about the former UK Prime Minister Margaret Thatcher using three born-digital source lists: curated website collections from the UK Web Archive, temporal search on Google (with query results set to past years) and a domain harvest of the .uk domain. The differences between the generated lists were rather striking: the corpus created by searching the UK Web Archive's collections was primarily governmental; the sources referring to Thatcher that appeared on Google with time stamps from the exact same period were primarily commercial (large newspaper and e-commerce platforms) and the corpus of the domain harvest was more diverse, containing a variety of types of sources.

This example shows that seed lists are embedded in specific web cultures, which may have internal preferences or biases. To understand these cultures and potential biases, in our case study of the reconstruction of the .yu domain from the Internet Archive, we used a fairly simple technique for comparing the seed lists that generated the reconstruction. This simple technique, initially coined by the Digital Methods Initiative as a triangulation tool[8], compared the source lists from which our reconstruction began (Google search results, Wikipedia, a Computer Magazine, and an ISP). First, we removed duplicates from each list, and then we identified the URLs shared among source types, and the URLs unique to each source type. The ties between the shared and unique sources can be read as a table, or visualized as a network graph. In the case of the .yu domain, we found that, relative to the other source lists, Wikipedia was the most diverse source to be used as a seed list for reconstructing topical corpora from the Internet Archive.

3. Assessing Incompleteness

Earlier in this chapter, we argued that an assessment of the (in)completeness of web archives is necessary for contextualizing historical research done with archived web materials. It should be noted that assessments of incompleteness greatly depend on metadata that is often unavailable or inaccessible for researchers, such as information about the collection policy, or access to the archiving crawler's logs.

Despite these limits, computational methods may give researchers the ability to assess the whole by comparing the archived dataset against different datasets and data sources, as was previously done by Thelwall and Vaughn (2004), Ainsworth et al. (2011), Alkawi et al. (2015) and Hale et al. (2017). In the case of the .yu domain, we assessed the number of *accessible*

archived snapshots of .yu websites on the Internet Archive, compared with the number of .yu URLs recovered by our outlink extraction method. Here, we queried the Internet Archive's Wayback Machine CDX server API[9] and documented the response code for each URL – indicating whether it can or cannot be accessed on the Wayback Machine. Using a python script and the urllib2 module, we sent GET queries to the REST API of the Internet Archive CDX server to lookup captures of our known Yugoslav URLs. With the help of filtering and collapsing features of the CDX server, we requested only successful snapshots and selected only one for each year. Using this method, we found that the level of archival coverage of the historical domain is a function of temporal proximity to the live web: the shorter the temporal gap between a live website and its archived snapshot, the better its archival coverage on the Internet Archive.

4. Culturomics: Histories of Non-Textual Elements

At the outset of this chapter we argued that some computational methods for textual analysis can be applied to any type of text, independently of their affiliation to various academic disciplines. This also applies to the archived web, where, for example, methods such as Topic Modeling have already been used to analyze the evolution of the textual content of historical websites over time (Milligan, 2016). At the same time, we also argued that the history of websites not only comprises text, but also other elements documenting the history of the medium, and the socio-technical 'ecology' in which a given website operated at the time it was archived (Berry, 2012; Fuller, 2005). We also mentioned the 'culturomic' approach (Michel et al., 2011), which calls for analyzing cultural patterns in large volumes of data. Although the archiving process does not and cannot capture the full ecosystem in which the live website operated, it still captures a wealth of data that can be of significant importance to web historians – especially those interested in technological and cultural histories. Apart from text, web archives contain images, video, various document formats such as MS Word, PDF or PowerPoint presentations; each HTML page also contains code, within which one can find historical 'treasures' such as embedded advertisements and cookies. Is it possible to conduct large-scale culturomics analysis of non-textual elements found in the Wayback Machine?

Inspired by Lev Manovich's work on cultural analytics (2009, 2012), we also argue that large-scale analyses of the non-textual elements of web archives may open up a variety of new historical analyses. To this end, we developed a computational tool for analyzing the evolution of color in non-photographic images of the reconstructed .yu domain, focusing on archived snapshots dated between 1997 and 2000 (Ben-David et al., 2018). We used a technique based on Machine Learning, to summarize the three dominant colors of each of the non-photographic

images in the reconstructed domain, in order to study the history of the ties between web design, cultural digital practices and preferences, and nationalism. Specifically, we compared the three-color histograms summarizing the images of the .yu domain with the colors of the Yugoslav flag, and calculated the overall 'distance' of the entire domain from the colors of the flag during the Kosovo war[10].

To summarize this section, it is evident that none of the computational techniques that we used for studying the history of the .yu domain are new. Rather, we adapted existing methods for web research and data analysis, and tailored them to meet the challenges of performing large-scale analysis of data with the Internet Archive's Wayback Machine. The following section concludes this chapter by discussing the limits of the computational approach to web historical research.

Conclusions

Although a computational approach brings with it a revolutionary promise of new paradigms of knowledge, and of writing and reading history, most of the advocates of this approach argue that it should be adopted judiciously – not necessarily as a replacement to existing humanist and social scientific methods, but rather as an important aid, a necessary instrument for gaining knowledge in a datafied world (Berry, 2012; Graham et al., 2015). In this chapter, we charted some of the new computational methods-based analysis strategies and approaches to studying archived websites accessed from the Wayback Machine. We have shown that despite the limits of the access interface – which is designed for viewing single archived pages rather than a distant reading of millions of them – python scripts can increase the analytical scope of historical web research, as well as answer important questions that improve the contextualization of archived web data (such as evaluation of sources and of archival coverage). We also demonstrated that computational methods can turn archived websites into a fascinating playground for identifying cultural trends by using the wealth of multimodal information they harbor.

Despite the benefits of the computational approach outlined above in stretching the analytical possibilities of historical research using the Wayback Machine, there are also several limits to consider. Chief of them is the production of analytical artifacts and their misinterpretation as historical facts. As previously mentioned, awareness of the cultures of sources and devices from which web archive research begins is crucial to the interpretation of the archival corpora and of the historical networks that each may produce. A possible way of overcoming the problem of analytical artifacts is triangulation of sources. In addition, analyses aided by computational methods, tools and techniques are able to answer specific quantifiable

questions. They are good at 'probing web archives' (Brügger, 2017) when dealing with huge corpora of archived websites, and when there is no alternative way of 'knowing' the archive as a whole. Nevertheless, this knowing is limited to quantifiable information, and macro analyses of millions of documents, or of petabytes of data, do not always suffice for answering deeper questions about historical processes (Winters, 2017). Further critical thinking and reflexive analysis are required to place the outputs of computational tools in context (Schafer et al., 2016).

References

Ainsworth, S.G., AlSum A., SalahEldeen H., Weigel, M.C., and Nelson, M. (2011) 'How much of the web is archived?' I: *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*. pp. 133–136.

Alkwai, L.M., Nelson, M.L., and Weigle, M.C. (2015) 'How well are Arabic websites archived?' In: *Proceedings of the 15th ACM/IEEE-CE Joint Conference on Digital Libraries*, 2015, pp. 223–232.

Baeza-Yates, R., Lalanne, F., Castillo, C., and Dupret, G. (2004) '*Comparing the Characteristics of the Korean and the Chilean Web*' *Korea-Chile IT Cooperation Center ITCC*. Technical Report, available at: http://chato.cl/ papers/baeza_04_comparing_chilean_web_korean_web.pdf

Ball, A. (2010) *DCC state of the art report: Web archiving*. http://www.dcc.ac.uk/sites/default/files/documents/reports/sarwa-v1.1.pdf

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 & Society*, 18(7): 1103–1119.

Ben-David, A. and Amram, A. (2018). 'The Internet Archive and the socio-technical construction of historical facts', *Internet Histories*, 2(1–2): 179–201.

Ben-David, A., Amram, A., and Bekkerman, R. (2018) 'The colors of the national web: Visual data analysis of the historical yugoslav web domain', *International Journal on Digital Libraries*, 19(1): 95–106.

Ben-David, A. and Huurdeman, H.C. (2014) 'Web archive search as research: Methodological and theoretical implications', *Alexandria*, 25(1–2): 93–111.

Berry, D. (2011) 'The computational turn: Thinking about the digital humanities', *Culture Machine*, 12: n.p, available at: http://www.culturemachine.net/index.php/cm/article/viewarticle/440

Berry, D. (2012) Life in Code and Software: Mediated Life in a Complex Computational Ecology. Open Humanities Press.

Bordino, I., Boldi, P., Donato, D. Santini, M., and Vigna, S. (2008) 'Temporal evolution of the UK Web', In: *Proceedings of the IEEE International Conference on Data Mining Workshops, ICDM Workshops*, 2008, pp. 909–918.

Brügger, N. (2012a) 'Web history and the Web as a historical source', *Zeithistorische Forschungen*, 9(2): 316–325.

Brügger, N. (2012b) 'When the present Web is later than the past: Web historiography, digital history, and Internet Studies', *Historical Social Research/Historische Sozialforschung*: 102–117. Available at: https://www.ssoar.info/ssoar/bitstream/handle/document/38378/ssoar-hsr-2012-4-brugger-When_the_present_web_is.pdf?sequence=1 (accessed 20 October 2015).

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

Brügger, N. (2016) 'Digital humanities in the 21st century: Digital material as a driving force', *DHQ: Digital Humanities Quarterly*: 10(3), available at: http://www.digitalhumanities.org/dhq/vol/10/3/000256/000256.html.

Brügger, N. (2017) 'Probing a nation's web domain: A new approach to web history and a new kind of historical source', In: Gerard Goggin and Mark McLelland (eds), *The Routledge Companion to Global Internet Histories*. New York: Routledge. pp. 61–73.

Chen, L. and Roy, A. (2009) 'Event detection from flickr data through wavelet-based spatial analysis', In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 2009, pp. 523–532.

Chung, Y.-J, Toyoda, M. and Kitsuregawa, M. (2009) 'A study of link farm distribution and evolution using a time series of web snapshots', In: *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web*, 2009, pp. 9–16.

Costa, M. and Silva, M.J. (2009) 'Towards information retrieval evaluation over web archives', In: *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, 2009, pp. 37–38.

Costa, M. and Silva, M.J. (2011) 'Characterizing search behavior in Web archives', In: *TWAW*, pp. 33–40.

Dougherty, M. Meyer, E.T., McCarthy Madsen, C., van den Heuvel, C., Thomas, A., and Wyatt, S. (2010) 'Researcher engagement with web archives: State of the art'. Joint Information Systems Committee Report.

Featherstone, M. (2006) 'Archive', *Theory, Culture & Society*, 23(2–3): 591–596. DOI: 10.1177/0263276406023002106.

Foot, K. and Schneider, S. (2010) 'Object-oriented web historiography', In: Niels Brügger (ed.), *Web History*. New York: Peter Lang. pp. 61–80. Available at: http://faculty.washington.edu/kfoot/Publications/Foot_Schneider.pdf (accessed 22 September 2017).

Foot, K., Schneider, S.M., Dougherty, M., Xenos, M., and Larsen, E. (2003) 'Analyzing linking practices: Candidate sites in the 2002 US electoral web sphere', *Journal of Computer-Mediated Communication* 8(4), available at: https://doi.org/10.1111/j.1083-6101.2003.tb00220.x.

Fuller, M. (2005) Media Ecologies: Materialist Energies in Art and Technoculture. London: MIT Press.

Graham, S., Milligan, I., and Weingart, S. (2015) *Exploring Big Historical Data: The Historian's Macroscope*. London: Imperial College Press.

Gyllstrom, K., Eickoff, C., de Vries, A.P., and Moens, M.F. (2012) 'The downside of markup: Examining the harmful effects of css and javascript on indexing today's web', In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 2012, pp. 1990–1994.

Hale, S.A., Blank, G., and Alexander, V.D. (2017) 'Live versus archive: Comparing a web archive to a population of web pages', In: Niels Brügger and Ralph Schroeder (eds), *Web as History: Using Web Archives to Understand the Past and the Present*. London: UCL Press. pp. 45–61. Available at: http://www.jstor.org/stable/j.ctt1mtz55k.8.

Hale, S.A, Yasseri, T., Cowls, J., Meyer, E.T., Schroeder, R., and Margetts, H.. (2014) 'Mapping the UK webspace: Fifteen years of British universities on the web', In: *Proceedings of the 2014 ACM Conference on Web Science*, 2014, pp. 62–70.

Holzmann, H., Goel, V., and Anand, A. (2016) 'Archivespark: Efficient web archive access, extraction and derivation', In: *Digital Libraries (JCDL), IEEE/ACM Joint Conference on*, 2016, pp. 83–92.

Holzmann, H., Nejdl, W., and Anand, A. (2017) 'Exploring Web archives through temporal anchor texts', In: *Proceedings of the 2017 ACM on Web Science Conference*, 2017, pp. 289–298.

Huurdeman, H.C., Ben-David, A., and Sammar, T. (2013) 'Sprint methods for web archive research', In: *Proceedings of the 5th Annual ACM Web Science Conference*, 2013, pp. 182–190.

Huurdeman H.C., Kamps, J., Samar, T., de Vries, A.P., Ben-David, A., and Rogers, R. (2015) 'Lost but not forgotten: Finding pages on the unarchived web', *International Journal on Digital Libraries* 16(3–4): 247–265. DOI: 10.1007/s00799-015-0153-3.

Jatowt, A., Kawai, Y., Ohshima, H., and Katsumi, T. (2008) 'What can history tell us?: Towards different models of interaction with document histories', In: *Proceedings of the 19th ACM Conference on Hypertext and Hypermedia*, 2008, pp. 5–14.

Koehler, W. (2002) 'Web page change and persistence – A four-year longitudinal study', *Journal of the Association for Information Science and Technology* 53(2): 162–171.

Lazer, D., Pentland, A. (Sandy), Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, M., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. (2009). Life in the network: the coming age of computational social science. Science (New York, N.Y.), 323(5915), 721–723. http://doi.org/10.1126/science.1167742 . (2009) 'Life in the network: The coming age of computational social science', *Science* 323(5915): 721–723. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/pmc2745217/ (accessed 22 September 2017).

Lin, J., Kraus, K., and Punzalan, R. (2014) 'Supporting "distant reading" for web archives', *Proceedings of Digital Humanities*: 239–241.

Lin, J., Milligan, I., Wiebe, J., and Zhou, A. (2017) 'Warcbase: Scalable analytics infrastructure for exploring Web archives', *Journal on Computing and Cultural Heritage* 10(4): 1–30.

Mataly, J. (2013) 'The Three Truths of Margaret Thatcher: Creating and Analyzing Archival Artefacts'. M.A dissertation, University of Amsterdam.

Manovich, L. (2009) The practice of everyday (media) life: From mass consumption to mass cultural production? *Critical Inquiry* 35(2): 319–331.

Manovich, L. (2012) How to compare one million images? In: David M. Berry (ed), *Understanding Digital Humanities*. London: Palgrave Macmillan. pp. 249–278.

Meyer, E.T., Thomas, A., and Schroeder, R. (2011) 'Web archives: The future (s)'. Available at SSRN: https://ssrn.com/abstract=1830025 or http://dx.doi.org/10.2139/ssrn.1830025

Michel, JB., SHEN, Y.K., Presser Aiden, A., Veres, A. Gray, M.K., The Google Books Team, Picket, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., and Liberman Aiden, E. (2011) 'Quantitative analysis of culture using millions of digitized books', *Science* 331(6014): 176–182. Available at: http://science.sciencemag.org/content/331/6014/176.short (accessed 22 September 2017).

Milligan, I. (2012) 'Mining the "Internet graveyard": Rethinking the historians' toolkit', *Journal of the Canadian Historical Association* 23(2): 21–64.

Milligan, I. (2016) 'Lost in the infinite archive: The promise and pitfalls of web archives', *International Journal of Humanities and Arts Computing* 10(1): 78–94.

Moretti, F. (2007) Graphs, Maps, Trees: Abstract Models for Literary History. New York: Verso.

Ras, M. and van Bussel, S. (2007) 'Web archiving user survey'. National Library of the Netherlands (Koninklijke Bibliotheek). Available at: https://www.kb.nl/sites/default/files/docs/kb_usersurvey_webarchive_en.pdf (accessed 22 September 2017).

Rauber, A. and Bina, H. (1999) 'A metaphor graphics based representation of digital libraries on the world wide web: Using the libviewer to make metadata visible', In: *Proceedings. Tenth International Workshop on Database and Expert Systems Applications*. IEEE, pp. 286–290.

Rauber, A. and Merkl, D. (1999) 'Mining text archives: Creating readable maps to structure and describe document collections', In: Jan M. Żytkow and Jan Rauch (eds), *Principles of Data Mining and Knowledge Discovery: Third European Conference, PKDD'99 Prague, Czech Republic*. Berlin: Springer. pp. 524–529.

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

Rogers, R. (2015) 'Digital methods for Web research', In: Robert A. Scott and Marlis C. Buchmann (eds), *Emerging Trends in the Social and Behavioral Sciences*. DOI: 10.1002/9781118900772.

Rogers, R. (2017) 'Doing Web history with the Internet Archive: Screencast documentaries', *Internet Histories* 1(160–172).

Schafer, V, Musiani, F., and Borelli, M. (2016) 'Web archiving, governance and STS'. *French Journal of Media Research*, 6, 1–23.

Schneider, S.M. and Foot, K.A. (2004) 'Web campaigning by US presidential primary candidates in 2000 and 2004', In: John C. Tedesco and Andrew Paul Williams (eds), *The Internet Election: Perspectives on the Web in Campaign*. Lanham: Rowman & Littlefield Publishers. pp. 21–36.

Schneider, S.M. and Foot, K.A. (2005) 'Web sphere analysis: An approach to studying online action', In: Christine Hine (ed.), *Virtual Methods: Issues in Social Research on the Internet*. New York: Berg. pp. 157–170.

Schostag, S. and Fønss-Jørgensen, E. (2012) 'Webarchiving: Legal deposit of internet in Denmark. A curatorial perspective', *Microform & Digitization Review* 41(3–4): 110–120.

Stirling, P., Chevallier, P., and Illien, G. (2012) 'Web archives for researchers: Representations, expectations and potential uses', *D-Lib* 18(3/4), available at: http://www.dlib.org/dlib/march12/stirling/03stirling.html.

Thelwall, M. and Vaughan, L. (2004) 'A fair history of the Web? Examining country balance in the Internet Archive', *Library & Information Science Research* 26(2): 162–176.

Voerman, G. (1998) 'Dutch political parties on the Internet', *ECPR News* 10(1): 8–9.

Weltevrede, E. and Helmond, A. (2012) 'Where do bloggers blog? Platform transitions within the historical Dutch blogosphere', *First Monday* 17(2). Available at: http://firstmonday.org/ojs/index.php/fm/article/view/3775/3142 (accessed 23 December 2015).

Winters, J. (2017) 'Breaking in to the mainstream: Demonstrating the value of internet (and web) histories', *Internet Histories* 1(1–2): 173–179. DOI: 10.1080/24701475.2017.1305713.

[1] Digital Methods Initiative, Wayback Machine Link Ripper, https://wiki.digitalmethods.net/Dmi/ToolInternetArchiveWaybackMachineLinkRipper (visited 01.02.18).

[2] Digital Methods Initiative, Internet Archive Wayback Machine Network per Year, https://wiki.digitalmethods.net/Dmi/ToolInternetArchiveWaybackMachineToNetwork (visited 01.02.18).

[3] There are several exceptions to this claim. The UK web archive is freely accessible online for the collections made before 2014, whereas the ccTLD archivings of the .uk domain (from 2014 onwards) can only be accessed onsite at one of five national libraries. As for the Danish web archive, researchers who receive permission can browse the web archive remotely, while MA students are allowed access to the web archive only in the reading rooms of the national library.

[4] Since 2016, this problem is slightly mitigated in the new interface of the Wayback Machine, which now provides information about the collection of web captures associated with the specific web crawl the capture came from. See https://blog.archive.org/2016/10/24/faqs-for-some-new-features-available-in-the-beta-wayback-machine/ (visited 01.02.18).

[5] See, for example, the Archives Unleashed Toolkit (Lin et al., 2017) and the ArchiveSpark project (Holzmann et al., 2016).

[6] Since 2016, the new version of the Wayback Machine allows for basic keyword search on phrases from the URL, yet full site search is not yet available. See https://blog.archive.org/2016/10/24/faqs-for-some-new-features-available-in-the-beta-wayback-machine/ (visited 01.02.18).

[7] Open Media and Information Lab, The Open University of Israel (2017). Internet Archive Link Extractor. https://github.com/omilab/internet-archive-link-extractor (visited 01.02.18).

[8] Digital Methods Initiative (2008). Triangulation. https://wiki.digitalmethods.net/Dmi/ToolTriangulation (visited 01.02.18).

[9] The Wayback CDX server API is a standalone HTTP servelet that serves the index that the Wayback Machine uses to lookup captures.

https://github.com/internetarchive/wayback/blob/master/wayback-cdx-server/README.md. (visited 01.02.18). For further research scenarios using the CDX file, see Milligan, 2016.

[10] Open Media and Information Lab, the Open University of Israel (2016). Image Color Analysis. https://github.com/omilab/image-color-analysis (accessed 22 September 2017).