MANCHESTER
1824
Manchester University Press

Web Archive Search as Research: Methodological and Theoretical Implications

ANAT BEN-DAVID AND HUGO HUURDEMAN

ABSTRACT

The field of web archiving is at a turning point. In the early years of web archiving, the single URL has been the dominant unit for preservation and access. Access tools such as the Internet Archive's Wayback Machine reflect this notion as they allowed consultation, or browsing, of one URL at a time. In recent years, however, the single URL approach to accessing web archives is being gradually replaced by search interfaces. This paper addresses the theoretical and methodological implications of the transition to search on web archive research. It introduces 'search as research' methods, practices already applied in studies of the live web, which can be repurposed and implemented for critically studying archived web data. Such methods open up a variety of analytical practices that were so far precluded by the single URL entry point to the web archive, such as the re-assemblage of existing collections around a theme or an event, the study of archival artefacts and scaling the unit of analysis from the single URL to the full archive, by generating aggregate views and summaries. The paper introduces examples to 'search as research' scenarios, which have been developed by the WebART project at the University of Amsterdam and the Centrum Wiskunde & Informatica, in collaboration with the National Library of the Netherlands. The paper concludes with a discussion of current and potential limitations of 'search as research' methods for studying web archives, and the ways with which they can be overcome in the near future.

Keywords: web archives, Internet Archive, Wayback Machine, search, national libraries

INTRODUCTION

Web archives are constantly challenged by the rapid changes of the medium. By the time web archiving practices consolidate, the web has already undergone significant changes, which require rethinking the archiving method once again. For example, current web archiving technologies, methods, file formats and best practices are well suited for archiving websites (Masanès, 2005), but are less suited for archiving new media forms and data formats such as social media platforms and applications for mobile phones.

http://dx.doi.org/10.7227/ALX.0022

WEB ARCHIVE SEARCH AS RESEARCH

Despite difficulties in keeping up with the emergence of new data formats, the practice of archiving websites has been consolidated and institutionalized. Since the establishment of the Internet Archive in 1996, web archiving technologies have been used to select, capture, archive and provide access to archives containing an ever-increasing number of websites. The significance of web archives for scholarly use was recognized early on, and research practices have emerged and evolved around the development of web archives.

In this paper we take interfaces to web archives as an object of study, through which both the history of the web, as well as that of the scholarly use of web archives may be read. As with archived websites, whose timestamps are determined by their crawl date, access interfaces to web archives have embedded temporalities which represent the web in the period during which they were conceived, and the types of imagined uses they privileged at that time. Since the early days of web archiving, the dominant interface to accessing web archives has been the Internet Archive's Wayback Machine, conceived in the late 1990s as a tool for browsing past versions of a single archived web page. In recent years, however, several web archives and research initiatives have started developing and implementing search interfaces to web archives. We describe the transition from the single URL access point to search interfaces as a turning point in the history of web archives, and discuss their methodological and theoretical implications for the scholarly use of web archives.

This paper is written from a researchers' perspective, and pays particular attention to the ties between researchers' needs, and the types of research that access interfaces to web archives privilege, or preclude. It starts by describing the early design of the Wayback Machine as a surfing and browsing interface, and the research practices that evolved around the single URL approach to accessing web archives. Subsequently, the paper discusses the transition to search interfaces, and introduces their potential methodological and theoretical implications for web archive research, by providing examples from the search interface developed by the WebART project for the Dutch National Library's web archive. The paper concludes by discussing the promise and limitations of search for advancing the scholarly use of web archives.

BROWSING THE WEB ARCHIVE

The Internet Archive was conceived at a time where the web was a relatively young medium, whose most outstanding innovative feature was hypertextuality. Through the early browsers, the web has been accessed from a URL, or a directory, and surfed through following links. Early search engines have attempted to organize the web through indexing and building directories, and the verbs most descriptive of its primary use were 'browsing' and 'surfing' (Rogers, 2013).

The establishment of the Internet Archive stemmed directly from the early search engine culture, and combined the logic of hypertextuality and archival principles to build the Internet Archive as a digital library. This perception is evident in Brewster Kahle's first piece on web archiving, published in *Scientific American* in 1997. Kahle, an information and computer scientist who founded the Internet Archive in 1996, describes how the web's hypertextual structure 'form(s) an informal citation system similar to the footnote system already in use. Studying the topography of these links and their evolution might provide insights into what any given community thought was important' (1997, np).

Kahle identifies a web archive with a web crawler, or a web indexer. Indeed, archiving, crawling and indexing were all suited to thinking about collection making, especially when the early search engines were perceived as the ultimate collection-makers of the web (Brin et al., 1998).

In 1996, the Internet Archive started to crawl and capture snapshots of websites. Five years later, after completing several web crawls, the archived pages were made available to the public through the Wayback Machine interface. The crawler of the Internet Archive, designed in tandem with Alexa search engine's technology for indexing the web, relies on the web's hypertextual structure, as it follows the links from one website to the other to discover new pages for capturing (Masanès, 2005; Rogers, 2013).

The Wayback Machine, accordingly, has been developed as an access tool to the Internet Archive's documents – the URLs. The envisioned use of the Internet Archive, through its interface, was that of browsing.[1] In a way, retrieving a single document from the archive for viewing and browsing not only reflects the early organizing culture of the web around hypertext, but also resembles the established practice of retrieving documents from an analogue archive, where the user would ask for a specific document to be called up from the depot, then sit at a table in the reading room and browse through the retrieved documents. The Wayback Machine is designed to retrieve a document by entering a URL, so that users can surf its archived snapshots at different points in time ('vertical surfing'), or follow links to other pages, if these have been archived ('horizontal surfing').

As Rogers notes, the initial perception of the Wayback Machine was to ensure the flow of surfing, since it served as a tool for consulting broken links that were found while surfing the live web (Rogers, 2013). This example

[1] As noted on the Frequently Asked Questions page of the Internet Archive: 'The Internet Archive Wayback Machine is a service that allows people to visit archived versions of Web sites. Visitors to the Wayback Machine can type in a URL, select a date range, and then begin surfing on an archived version of the Web. Imagine surfing circa 1999 and looking at all the Y2K hype, or revisiting an older version of your favorite Web site. The Internet Archive Wayback Machine can make all of this possible'. URL https://web.archive.org/web/20140312062753/http://archive.org/about/faqs.php (visited 25.4.14).

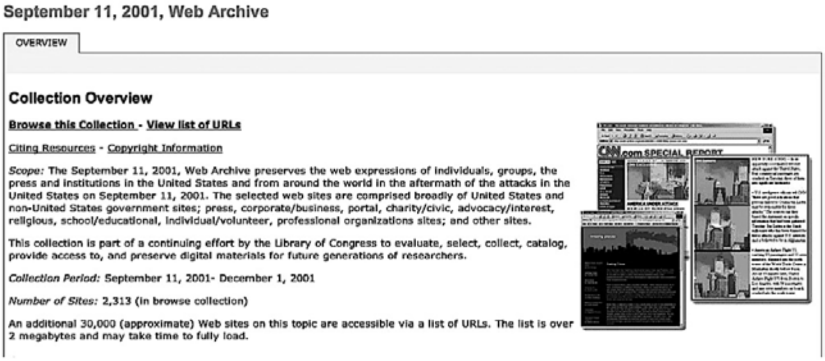
shows the extent to which the web archive was perceived in its early years as a good-enough copy of the world wide web. Considering the size of the web in 1997, the 1997 crawl of the Internet Archive may have been the most comprehensive archiving crawl of the web (Soboroff, 2002).

EARLY SCHOLARLY USES OF WEB ARCHIVES

Along with the development of the Wayback Machine, and especially after the rapid establishment of web archiving initiatives at national libraries around the world, web archive research practices have emerged. Since the early 2000s, a gradually evolving scholarship has begun referring to web archives both as a tool for historiographical research, as well as an object of study in itself (Arms et al., 2006; Ball, 2010; Brügger, 2009; Dougherty et al., 2010; Thomas et al., 2010). The Internet Archive has been used as a source for performing longitudinal analysis of the evolution of hyperlinks over time (Kraft et al., 2003), evolution of E-commerce websites (Chu et al., 2007), accessibility of higher education websites (Hackett and Parmanto, 2005) and of link structures of academic websites (Björneborn, 2004; Vaughan and Thelwall, 2003). Later studies have also examined specific themes using data from the Internet Archive, such as website quality of the UK airline industry (Xie and Barnes, 2008), availability and persistence of web citations (Casserly and Bird, 2008), or the development of the concept of sharing in blogs (John, 2013). Legal scholars have also discussed the consequences and possibilities of using web archives as evidence (Howell, 2006; Eltgroth, 2009; Rogers, 2013).

In addition, website historiography has emerged as a new genre through which the evolution of a website's content, design or other technological elements are analysed and traced (Brügger, 2010; Rogers, 2013). Apart from the single website historiography, the Wayback Machine has also been used as a tool for studying histories of other media such as the history of online newspapers, broadcasting media or the history of software (Ankerson, 2012; Falkenberg, 2010).

The early work of Foot and Schneider on web sphere analysis, in particular, has established a research practice that involves a dynamic selection and archiving of a set of webpages around a theme or an event, which are subsequently analysed using a triangulation of hyperlink, content and qualitative analyses (Schneider and Foot, 2004a; Foot and Schneider, 2010; Foot et al., 2003). Web sphere analysis has been used by social scientists for studying and comparing electoral campaigns over time, while taking into account both the content, hyperlink structure and other digital elements of candidates and voters' websites (Foot et al., 2007; Jankowski et al., 2005; Kluver, 2007). Web sphere analysis has also been used for documenting and analysing disasters, whether man-made or natural. For example, the 2001 September 11 web archive, curated by Foot and Schneider (Foot et al., 2005), has been studied for analysing personal expressions of bereave-



September 11, 2001, Web Archive

Collection Overview

Browse this Collection - View list of URLs

Citing Resources - Copyright Information

Scope: The September 11, 2001, Web Archive preserves the web expressions of individuals, groups, the press and institutions in the United States and from around the world in the aftermath of the attacks in the United States on September 11, 2001. The selected web sites are comprised broadly of United States and non-United States government sites; press, corporate/business, portal, charity/civic, advocacy/interest, religious, school/educational, individual/volunteer, professional organizations sites; and other sites.

This collection is part of a continuing effort by the Library of Congress to evaluate, select, collect, catalog, provide access to, and preserve digital materials for future generations of researchers.

Collection Period: September 11, 2001- December 1, 2001

Number of Sites: 2,313 (in browse collection)

An additional 30,000 (approximate) Web sites on this topic are accessible via a list of URLs. The list is over 2 megabytes and may take time to fully load.

Figure 1: A screenshot from the Library of Congress September 11 Web Archive, http://lcweb2.loc.gov/diglib/lcwa/html/sept11/sept11-overview.html, taken in September 2013

ment (Siegl and Foot, 2004) and for conceptualizing online structures of civic engagement at times of crisis (Schneider and Foot, 2004b). It should be noted, however, that the collection of archived websites still assumes a practice of viewing single pages, or browsing through the collection[2] (see Figure 1).

In studying the types of historiographical research privileged by web archives, Richard Rogers and researchers from the Digital Methods Initiative have built tools on top of the Wayback Machine to repurpose its output for social research. The Wayback Machine Link Ripper tool,[3] for example, outputs a list of direct links to all archived versions of a given URL, which can then be used to facilitate the narration of a single site historiography, or the curating of a collection of archived websites. The 'Internet Archive Wayback Machine Network Per Year' tool extends the scope of analysis beyond the single URL by allowing researchers to insert a set of URLs.[4] The tool subsequently retrieves the archived versions of each URL closest to 1 July for a specific year, and extracts a hyperlink network as its output. Such tools have been used to study the evolution of hyperlink networks over time, such as the Dutch blogosphere (Weltevrede and Helmond, 2012).

[2] Search options were added to the Library of Congress' Web Archive between January and April 2014, see https://web.archive.org/web/20140407141559/http://lcweb2.loc.gov/diglib/lcwa/html/sept11/sept11-overview.html (visited 25.4.14).

[3] Digital Methods Initiative, *Wayback Machine Link Ripper*, https://wiki.digitalmethods.net/Dmi/ToolInternetArchiveWaybackMachineLinkRipper (visited 25.4.14).

[4] Digital Methods Initiative, *Internet Archive Wayback Machine Network per Year,* https://wiki.digitalmethods.net/Dmi/ToolInternetArchiveWaybackMachineToNetwork (visited 25.4.14).

Types of research precluded by the single-document approach

Despite the varied research practices described above, the single URL approach proposes obstacles to the scholarly use of web archives. Arguably, access to web archives has remained tied to the web's early user engagement practices (surfing, browsing), and did not keep up with the search culture and practice that evolved with the web. Put differently, most web archives are not searchable.

To consult the web archive through the Wayback Machine, a researcher must know the URL she would like to retrieve from the archive. Knowing the URL stems from the notion of consulting the Internet Archive while browsing the live web (and finding a broken link, or a URL whose past versions are interesting to view), assuming a temporal proximity between the live and archived objects. However, as the temporal distance between the disappearance of a page from the live web and its archived version increases, knowing what would be a relevant URL of the past to type into the interface would be much more difficult, if not impossible. Thus without the ability to search for keywords or other contextual elements, it would be difficult for future historians to trace the relevant URL as the starting point to studying a theme, an issue or an event using web archives.

In addition, even when the URLs of interest are known, the researcher does not and cannot know whether the archive contains other URLs that may be of relevance. In many cases, researchers use external sources, or 'expert lists' from which the list of URLs is retrieved (Weltevrede and Helmond, 2012). Without the external source for URLs, building thematic collections of archived pages based on viewing single documents requires a manual and time-consuming process of browsing through URLs that serve as starting points, checking through their past versions and the links from each page to find relevant documents to be included in the collection. Apart from the laborious manual harvesting and calculation of hyperlink networks of archived pages, there remains a question of the documents' timestamps: The timestamp of the snapshot of a past version of a URL is that of the date of archiving, not necessarily the last updated date of that URL. As Rogers argues, the Wayback Machine 'jump cuts through time', since it retrieves the archived pages whose crawled date is closest to the crawl date of the previously browsed page. Horizontal browsing of the archive, then, is not necessarily horizontal, as the timestamps of the retrieved pages are not identical, and collecting archived pages from the same crawl date for reconstructing past hyperlink networks is a challenging task. To solve this problem, researchers usually aggregate the archived URLs per year, which results in an approximation of an historical hyperlink network with a large margin of error.

The single URL approach to accessing web archives and performing research with them has dominated the field in its constitutive years. In recent years, web archiving initiatives have begun developing new access interfaces to web archives, to advance their scholarly use. Several leading national web archives, such as the Japanese, British and Portuguese, already offer full-text search as an alternative access point to the archived pages, and other web archiving institutions are currently working on adapting their infrastructure and backend to support search interfaces in the near future (Costa and Silva, 2011; Gomes et al., 2011). The following section of this article discusses the methodological and theoretical implications of the introduction of search interfaces to the scholarly use of web archives.

SEARCHABLE WEB ARCHIVES: A TURNING POINT IN WEB ARCHIVE RESEARCH

The continuous growth in size and volume of web archives presents infrastructural challenges to web archiving institutions. Such challenges also apply to building search interfaces on top of existing web archives, whether small or large. To build search interfaces on top of web archives, the content of the entire archive must first be indexed. While indexing does not fundamentally change the organizing principle of web archives that are still structured around separate ARC or WARC files for each crawled website, it introduces the full archive as unit of analysis, as indexing entails knowing what the entire archive contains. Moreover, the index of a web archive weaves the archive's data in ways that both break the boundaries between single files, as well as within each file (the non-textual elements, such as anchor text, images, timestamps, crawl dates). Indexing therefore introduces a new retrieval model for the web archive: instead of retrieving a single document contained in a file, the indexed web archive becomes a dataset. Querying the index could therefore be seen as an act of reassembling, where documents and other elements from the archive are retrieved and reorganized in real time to match the query's setting.

It should be noted, however, that full-text search interfaces to web archives do not replace the Wayback Machine, which continues to function as the primary tool used by most web archiving initiatives for viewing the archived pages. Instead, search interfaces add new ways of exploring and engaging with the archive's content, before specific pages are viewed. First, full-text search interfaces allow researchers to retrieve all the archive's documents that contain the relevant keywords of interest, which also facilitates the creation of thematic collections. (In a sense, full-text search results are in themselves a thematic collection.) Second, search allows researchers to move beyond the practice of building collections, to comparing and critiquing them. Third, researchers may be able to scale the unit of analysis, based on their research question. A historian studying an event, or a specific period, for example, may be interested in a delineated collection of websites; a computational linguist, on the other hand, may want to study the full archive as a corpus for analysis. Fourth, as previously mentioned, indexing the archive also enables adding metadata from the archive's backend to the search results. This may enable accessing the archive not just through typing a URL or performing textual search, but also by retrieving other non-textual components, such as

Table 1: URL-based and search-based access to web archives compared

	Single URL	Search
Unit of analysis	Webpage	Archived web data
Wayback Machine	Primary analytical tool	Viewing/validation tool
Focus (content)	Text	Digital objects (text, hyperlinks, metadata, images)
Collections	Building (manually)	Reassembling/critique

time ranges and geographic location. Combined, these possibilities constitute new access and retrieval options for web archives. The primary differences between the single URL and search access points to web archives are summarized in Table 1.

In the same way that the single URL approach privileges specific types of web archive research, searchable web archives also have methodological and theoretical implications, which are discussed in the following section of this article.

Exploring new research scenarios for web archive research

In the past decade, search has become the dominant paradigm for user engagement with the web. A study on information needs of users of web archives indicates that web users are so accustomed to seeking information through querying search engines that they also expect to find information in web archives in the same way (Costa and Silva, 2011). To meet researchers' needs, several national web archives already offer full-text search interfaces, and many other web archiving initiatives are working on the implementation of full-text search in the near future.

However, user engagement (either with the live web or a web archive) is not necessarily a research practice. Rogers has put forward the notion of 'search as research', to refer to a set of practices that repurpose search from an information-seeking tool into a method for the social study of the web (Rogers, 2013). Search as research is aimed in particular at studying search engine algorithms as a device that organizes and ranks the web according to parameters that are in most cases hidden from the users (for example, the size of a search engine's index, or the influence of the number of incoming links, geolocation of IP address, or a user's browsing history on the ranking of search results). Through query design and comparative analysis of search results, search as research aims at making search engine's algorithmic authority transparent. At the same time, since search engines do organize knowledge on the web, their algorithmic authority can be used for social studies of the web, for example by studying the standing of an issue by looking at the 'distance' of the ranking of results of a specific query from the top of the list (source-distance analysis). In a similar fashion, the commitment of specific actors to an issue may be read by the list of actors populated by search results of specific queries (issue commitment analysis).

As the following section of the paper shows, query design, source-distance analysis or issue-commitment analysis can be readily applied as a research approach to searchable web archives. Examples are given from WebARTist, a prototype search interface of the Dutch Web Archive at the National Library of the Netherlands, which was developed by researchers from the University of Amsterdam and the Centrum voor Wiskunde en Informatica (CWI).

WebARTist as a research engine

The Web Archiving Retrieval Tools (WebART) project at the University of Amsterdam and the CWI aims to critically assess the value of web archives for realistic research scenarios, and to develop novel information access tools and methods to maximize the archive's utility for research.[5] WebART uses a 'Living Lab' setting, in which researchers from different disciplines cooperatively explore ways to implement and refine access tools for web archives. Our approach is to conduct actual web archive research hand-in-hand with the development of web archive access tools tailored to realistic research scenarios.

During the pilot year of the project, we have developed WebARTist (Web Archive Temporal Information Search Tools), a dynamic, full-text search engine of the Dutch Web Archive, which supports flexible options for visualizations, exports, aggregation and filtering of search results. The interface allows the users to formulate a query, and view the list of results that are ranked based on standard information retrieval models[6] (Huurdeman et al., 2013). Next to the title of the URL containing the search word, the search results also include the snippet text from the body of the retrieved page, as well as other metadata such as the crawl date of the page, the last updated date of the page, the collection in which it is archived and its size. Eventually, the user may view the URL through the Dutch Web Archive's Wayback Machine interface, as well as through a link to the Internet Archive's Wayback Machine (see Figure 2).

WebARTist also allows for filtering the search results based on various criteria, such as a specific time range, a specific crawl, a collection, or a website. Finally, WebARTist allows users to export the results to formats suitable for further processing (Huurdeman et al., 2013).

Using the interface, one may engage in a critical study of the Dutch Web Archive, for example, by comparing the results of different queries. Historians interested in studying the 2008 economic crisis in Europe, for example, may query the word 'Eurocrisis' and generate a tag cloud from the

[5] WebART, http://www.webarchiving.nl (visited 25.4.14).

[6] The default retrieval model of WebARTist is BM25, although it is possible to select other retrieval models in the interface.

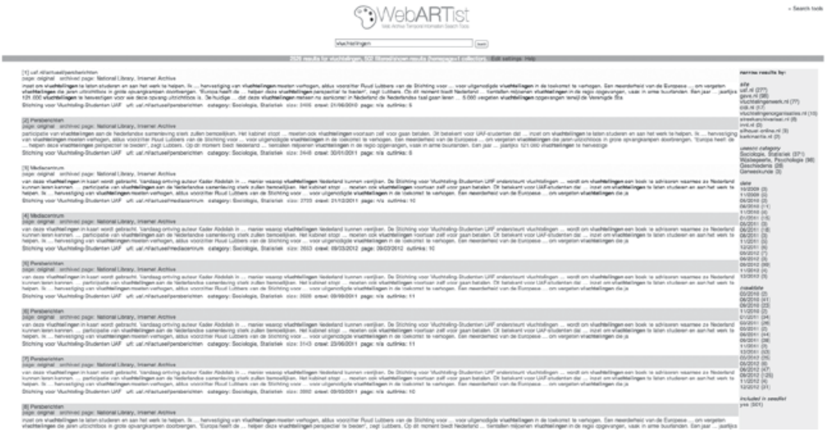


Figure 2: A screenshot of the prototype of the WebARTist search system

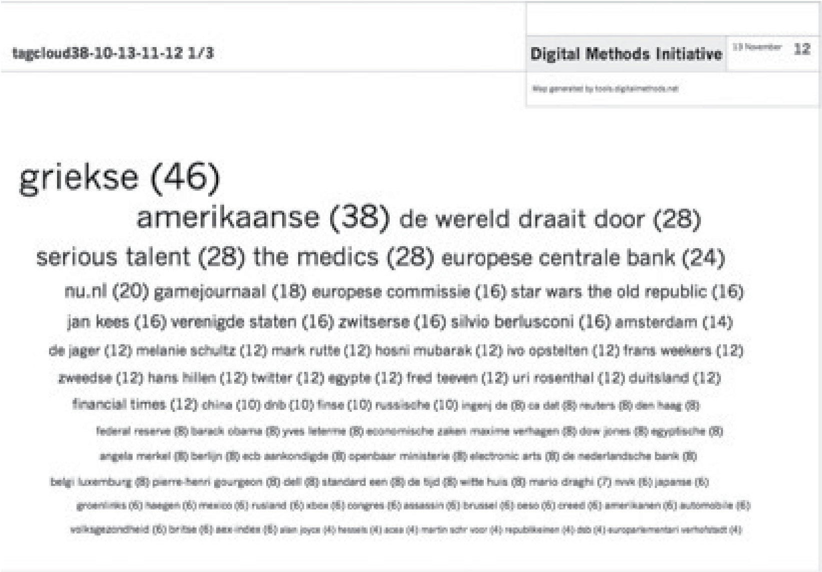
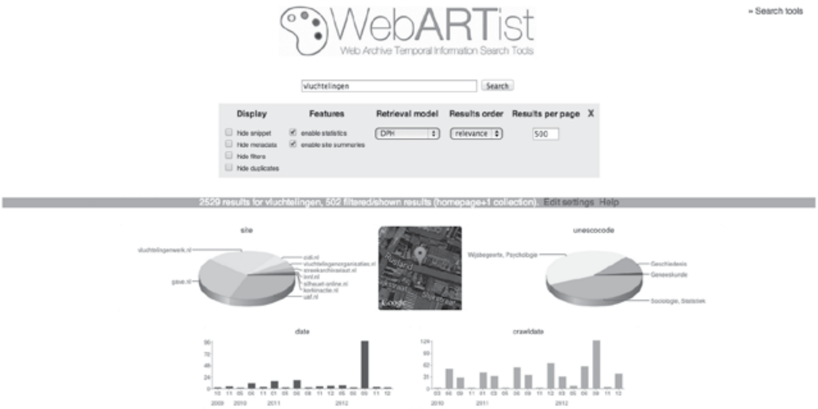


Figure 3: A tag cloud consisting of the snippet text of all results returned by WebARTist to the query 'eurocrisis' in November 2012

snippet text of the returned results to see which countries and issues were most associated with the crisis at a specific point in time. As seen in Figure 3, the tag cloud suggests that Greece, the United States and the European Central Bank are the most mentioned actors in the Dutch Web Archive in this context.

WebARTists's features make use of additional information found at the back-end of the archive to enrich the research interface. In one prototype of the search interface, for example, an aggregation of the search results generates statistics about the relative standing of the search results compared to the entire archive across different parameters, such as the distribution of websites, sub-collections, and crawl dates. For example, the aggregate results of the query 'vluchtelingen' (refugees), show that the issue space is populated by nine websites, among which www.vluchtelingenwerk.nl, a website of an organization dedicated to assisting refugees through the asylum procedure, has the most mentions of the term. The aggregated statistics also show the distribution of thematic categories assigned to websites archived by the collection department of the National Library of the Netherlands. In the example of the query 'vluchtelingen', 40% of the results have been categorized under 'Sociology and Statistics', another 40% under the category 'Philosophy and Psychology', and the remaining 20% have been categorized under 'History' and 'Medicine' (see Figure 4). Other data enrichments at the archive's backend include the building of a geo-index, based on querying the entire archive's content for all Dutch street names and postcodes, an image index and a calculation of the archive's web graph, which allows users to export the outlinks found in each search result to facilitate hyperlink analysis (see Figures 4–5).

In a sense, the aggregate views of the entire web archive also mark the shift from the focus on the single archived URL to the entire web archive as unit of analysis. The aggregated statistics also give rise to methods that rescale the unit of analysis by reassembling collections or building subsets. One may, for example, slice the entire archive by a specific time range, or a specific crawl, or create a sample of the top 5,000 interlinked pages in the archive.



vluchtelingen

Display	Features	Retrieval model	Results order	Results per page
□ hide snippet	☑ enable statistics	DPH	relevance	500
□ hide metadata	☑ enable site summaries			
□ hide filters				
□ hide duplicates				

2529 results for vluchtelingen, 502 filtered/shown results (homepage+1 collection). Edit settings Help

Figure 4: A screenshot from WebARTist displaying the aggregate statistics for the query 'vluchtelingen' (refugees)

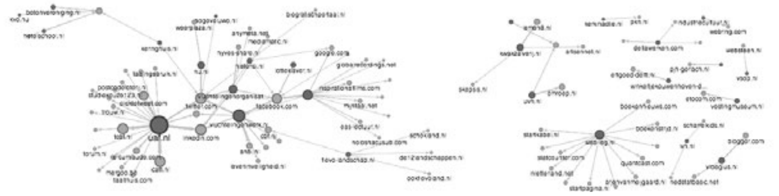
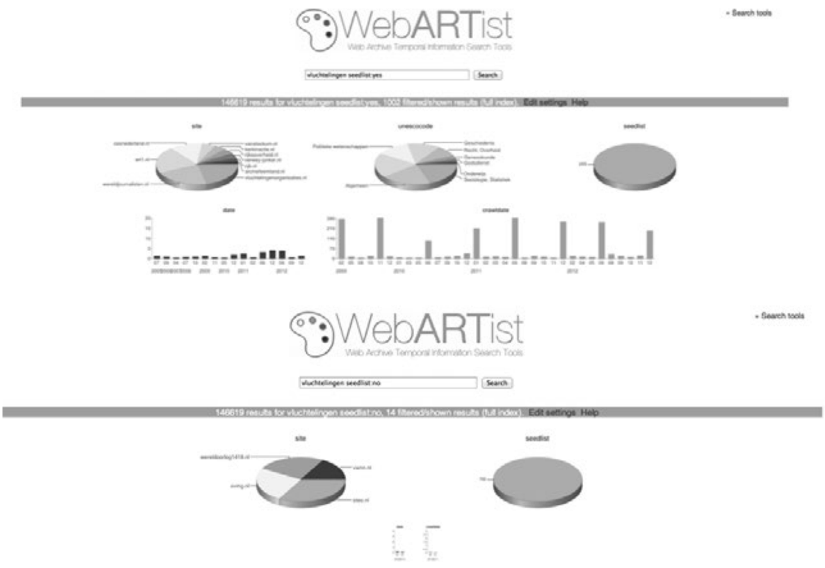


Figure 5: A hyperlink network generated from the automatic extraction of outlinks from the search results to the query 'vluchtelingen' in the Dutch Web Archive. Visualization by Google Fusion Tables

Moreover, treating the full archive as a unit of analysis also gives rise to practices of archival critique. Through analysing the archive's selection policy, interlinked structure and development over time, one may, for example, identify 'archival artefacts', that is, phenomena or objects that are an outcome of the archiving method or selection policy. In a preliminary study of the Dutch Web Archive, which has a selective policy for archiving a pre-selected list of about 5,000 websites, we found that almost 10% of the pages contained in the archive were not part of the seed list, but were automatically captured due to the crawlers' setting (Samar et al., 2014). Subsequently, the distinction between pages that were part of the seed list and pages that were unintentionally archived were added to WebARTist as a feature that allows for greater transparency and contextualization of the search results, as well as for a critical comparison between the standing of an issue in its intentionally and unintentionally archived environments. In the case of the query 'vluchtelingen' mentioned above, for example, there are additional fourteen unintentionally archived pages, a website dedicated to the commemoration of the Great War among them, which may be relevant to studying the issue space. Alternatively, this method may also be of use to the curators of the web archive as a means for discovering new 'candidates' for archiving (see Figure 6).

In the same way that the Wayback Machine reflects early user engagement practices of the web, the scaling of web archives' unit of analysis from the single document to the entire archive is also a reflection of emerging practices in Digital Humanities and big data analytics, which propose methods for analysing very large datasets and digital collections, such as the use of N-gram search for analysing the temporal evolution of specific terms in large collections containing millions of digitized books (Michel et al., 2011).[7]

[7] The UK Web Archive has recently introduced n-gram searh. See http://www.webarchive.org.uk/ukwa/ngram/ (visited 25.4.14).



vluchtelingen seedlist:yes

146619 results for vluchtelingen seedlist:yes, 1002 filtered/shown results (full index).

vluchtelingen seedlist:no

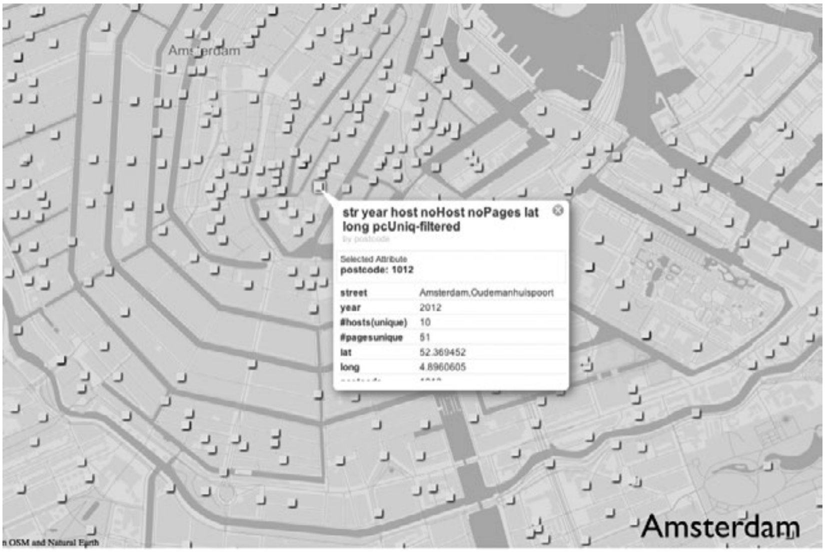
146619 results for vluchtelingen seedlist:no, 14 filtered/shown results (full index).

Figure 6: A screenshot from WebARTist comparing the aggregate distribution of archived websites that contain the word 'vluchtelingen', between seed-list and non seed-list pages in the Dutch Web Archive

The data-driven approach that breaks away from the simple semantic search of text, may characterize the next turn in web archiving research. Future interfaces to web archives may also reuse features extracted from the entire archive's data to develop alternative access points to the web archive. The following hypothetical example proposes a location-based access point to the web archive, in which users would be able to click on an area in a geographical map, or enter a postal code and a time range, to study the web history related to a specific place or location, or to produce geographical 'heat maps' that show the intensity of mentions of place names associated with specific issues, and their evolution over time (see Figures 7–8).[8]

Since in many national libraries online access to web archives is restricted by copyright and privacy issues, aggregate views of web archives may also function as an alternative interface that may already be implemented to grant users and researchers the ability to access and analyse the web archive, without accessing or viewing the archived pages.

[8] The UK Web Archive was the first to create a geo-index of the JISC UK Web Domain Dataset (1996–2010), which is available for download. See http://www.webarchive.org.uk/ukwa/visualisation/ukwa.ds.2/geo (visited 25.4.14).



str year host noHost noPages lat long pcUniq-filtered

by postcode

Selected Attribute
postcode: 1012

street	Amsterdam,Oudemanhuispoort
year	2012
#hosts(unique)	10
#pagesunique	51
lat	52.369452
long	4.8960605

Figure 7: A screenshot from a hypothetical postcode entry point to the Dutch Web Archive. The map shows that in 2012 there were ten unique hosts (websites) and fifty-one unique webpages that contained the postal code of the University of Amsterdam

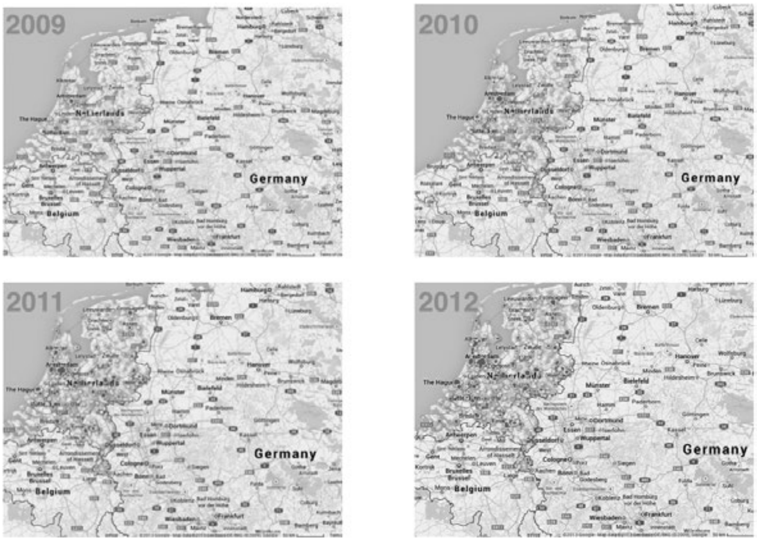


Figure 8: A geographical heat map, displaying the increase in the mentions and distribution of place names in the Dutch Web Archive between 2009 and 2012

CONCLUSIONS: BEYOND SEARCH

This article charts a turning point in the field of web archiving and web archive research, by analysing the methodological and theoretical implications of the transition from the single URL as the dominant unit of access and analysis, to search interfaces. This transition is also reflective of the evolution of user engagement practices with the live web, from browsing, to searching, and – more recently – to big data analytics.

The examples from this article show that searchable web archives lend themselves to a variety of research scenarios, such as the comparative analysis of search results, the study of the standing of issues or topics in particular points in time, the identification of archival artefacts, or the re-assemblage of existing collections based on thematic, temporal or technical criteria. Searchable web archives also introduce the entire archive as unit of analysis, which may be analysed using methods that borrow from digital humanities and big-data analytics, such as N-gram search and web graph analysis.

However, like the Wayback Machine, which lends itself to particular types of analyses and precludes others, search interfaces to web archives also have limitations, which present new methodological and theoretical challenges that are yet to be resolved. Some of these challenges relate to the interpretation of search results.

When researching the live web and comparing the number of search results, or the ranking of websites returned to competing queries by the same search engine, one can infer the standing of an issue on the web, according to the search engine's algorithmic authority (Rogers, 2013). In the case of web archive search, however, the number of returned results indicates the absolute number of pages *contained in the archive*, but lacks significant indication to their standing in the live web at the time of archiving. For example, when comparing the search results of two presidential candidates, the fact that the web archive contains more mentions of one candidate than another does not indicate that this candidate was more popular on the web at the time of archiving, since the number of mentions in the archive may be determined by the archive's selection policy, or influenced by the specific timing of the crawl. The number of mentions may also include duplicates of the same page, which were captured at different points in time, or influenced by the number of pages that each website contains. Moreover, the use of 'syntactic matching' in retrieval models may include 'false positives' in the search results. To name one anecdotal example, a search in the Dutch web archive for the Queen's name 'Maxima', returns many pages reporting the weather in the Netherlands, since 'Maxima' is also the Dutch word that describes the day's highest temperature. In other words, without supporting contextualization, search alone does not suffice as an analytical method for web archive research.

In addition, there are two significant differences between searchable web archives and search engines of the live web that researchers should be aware of. The first concerns the completeness of the index, and the second concerns temporality. As previously mentioned, the size of commercial web search engines' index is never made public, and users have to 'trust' the algorithm when interpreting the number of search results found. Put differently, the displayed number of results is interpreted as an estimate, while the displayed number of search results in a web archive is absolute, since the size of the full index is known. The second difference concerns the fact that commercial web search engines prioritize the ranking of fresh and recently updated websites as more relevant than older, static websites. Since most search engines do not display results beyond the 1,000 ranking, older websites often do not appear in search results. By contrast, temporality does not influence the ranking of web archive search results, which display both old and recent archived documents based on their semantic relevance. Researchers unaware of these differences may mistakenly confuse the ranking algorithm of the web archive with that of search engines of the live web, which might lead to a misinterpretation of search results.

It is important to mention that to date, the Wayback Machine is still the most prevalent interface to most web archives, and that most search interfaces to web archives have not yet been fully implemented or made available for scholarly use. It is also important to note that to date, web archives are not widely used for research (Meyer et al., 2011). In 2010, researchers from the Oxford Internet Institute in the United Kingdom and the Virtual Knowledge Studio in the Netherlands issued two expert reports titled 'researcher engagement with web archives'. Both reports mention a gap between the efforts of large libraries to build large multipurpose web archives, and between the researchers' needs for specific types of analyses (Dougherty et al., 2010; Thomas et al., 2010). Future research may be able to analyse retrospectively whether the introduction of search interfaces has significantly changed, or contributed to advancing researchers' engagement with web archives.

REFERENCES

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

Arms, William Y., Aya, Seluck, Dmitriev, Pavel, Kot, Blazej, Mitchell, Ruth and Walle, Lucia (2006) 'A research library based on the historical collections of the Internet Archive'. *D-Lib Magazine*, 12(2), http://www.dlib.org/dlib/february06/arms/02arms.html (visited 25.4.14).

Ball, Alex (2010) 'DCC state of the art report: web archiving', https://www.era.lib.ed.ac.uk/bitstream/1842/3327/1/Ball%20sarwa-v1.1.pdf (visited 25.4.14).

Björneborn, Lennart (2004) *Small-world link structures across an academic web space: a library and information science approach*. Copenhagen Royal School of Information and Library Science.

Brin, Sergey, Motwani, Rajeev, Page, Lawrence and Winograd, Terry (1998) 'What can you do with a web in your pocket?' *Data Engineering Bulletin,* 21, pp. 37–47, DOI 10.1.1.36.2806.

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

Brügger, Niels (2010) *Web history.* New York: Peter Lang.

Casserly, Mary. F. and Bird, James E. (2008) Web citation availability – a follow-up study. *Library Resources & Technical Services*, 52(1), pp. 42–53.

Chu, Sung-Chi, Leung, Lawrence, Hui, Yer Van and Cheung, Waiman (2007) 'Evolution of e-commerce Web sites: A conceptual framework and a longitudinal study'. *Information and Management,* 44(2), pp. 154–164.

Costa, Miguel and Silva, Mário J (2010) 'Understanding the information needs of web archive users'. *In Proceedings of the 10th International Web Archiving Workshop*, pp. 9–16, DOI: 10.1.1.368.3809.

Costa, Miguel and Silva, Mário J (2011) 'Characterizing search behavior in web archives'. In *Proceedings of the 1st International Temporal Web Analytics Workshop*, DOI: 10.1.1.368.5381.

Dougherty, Meghan, Meyer, Eric T., Madsen, Christine McCarthy, van den Heuvel, Charles, Thomas, Arthur and Wyatt, Sally (2010) 'Researcher engagement with web archives: state of the art'. *Joint Information Systems Committee Report*, first published online 1 Aug 10, http://ssrn.com/abstract=1714997.

Eltgroth, Deborah R. (2009) 'Best evidence and the Wayback Machine: toward a workable authentication standard for archived Internet evidence'. *Fordham Law Review,* 78(5), p. 181, first published online 22 Aug 2009, http://ssrn.com/abstract=1459805.

Falkenberg, Vidar (2010) '(R)evolution under construction: the dual history of online newspapers and newspapers online'. In: Brügger, Niels (ed.), *Web history*. New York: Peter Lang, pp. 233–256.

Foot, Kirsten and Schneider, Steven (2010) 'Object-oriented web historiography'. In: Brügger, Niels (ed.), *Web history*. New York: Peter Lang, pp. 61–79.

Foot, Kirsten, Schneider, Steven, Dougherty, Meghan, Xenos, Michael and Larsen, Ellena (2003) 'Analyzing linking practices: candidate sites in the 2002 US electoral web sphere'. *Journal of Computer-Mediated Communication 8(4)*, DOI: 10.1111/j.1083–6101.2003.tb00220.x.

Foot, Kirsten, Warnick, Barbara and Schneider, Steven (2005) Web-based memorializing after September 11: toward a conceptual framework. *Journal of Computer-Mediated Communication,* 11(1), pp. 72–96.

Foot, Kirsten, Schneider, Steven, Kluver, Randolph, Xenos, Michael and Jankowski, Nicholas (2007) 'Comparing web production practices across electoral web spheres'. In: Kluver, Randolph, Jankowski, Nicholas, Foot, Kirsten and Schneider, Steven (eds), *The internet and national elections:*

a comparative study of web campaigning. New York: Routledge, pp. 243–259.

Gomes, Daniel, João, Miranda and Costa, Miguel (2011) 'A survey on web archiving initiatives'. *Research and advanced technology for digital libraries*. Berlin Heidelberg: Springer, pp. 408–420.

Hackett, Stephanie and Bambang Parmanto (2005) 'A longitudinal evaluation of accessibility: higher education web sites'. *Internet Research,* 15(3), pp. 281–294.

Howell, Beryl (2006) 'Proving web history: How to use the Internet Archive'. *Journal of Internet Law,* 9(8), pp. 3–9.

Huurdeman, Hugo, Ben-David, Anat and Samar, Thaer (2013) 'Sprint methods for web archive research'. In *Proceedings of the 5th Annual ACM Web Science Conference*. New York: ACM, pp. 182–190, DOI: 10.1145/2464464.2464513.

Jankowski, Nicholas, Foot, Kirsten, Kluver, Randolph and Schneider, Steven (2005) 'The web and the 2004 EP election: Comparing political actor web sites in 11 EU Member States'. *Information Polity*, 10, pp. 165–176.

John, Nicholas. A. (2013) Sharing and Web 2.0: the emergence of a keyword. *New Media & Society*, 15(2), pp. 167–182.

Kahle, Brewster (1997) 'Preserving the internet'. *Scientific American,* 276, pp. 82–83.

Kluver, Randolph (ed). (2007) *The internet and national elections: comparative study of web campaigning*. Vol 2. London: Taylor & Francis.

Kraft, R., Hastor, E. and Stata, R. (2003) TimeLinks: Exploring the link structure of the evolving Web. In *WAW2003 Second Workshop on Algorithms and Models for the Web-Graph*.

Masanès, Julien (2005) 'Web archiving methods and approaches: a comparative study'. *Library Trends,* 54(1), pp. 72–90.

Meyer, Eric, Thomas, Arthur and Schroeder, Ralph (2011) 'Web archives – the future(s)'. *Oxford Internet Institute / IIPC*, first published online 30 June 11, DOI: 10.2139/ssrn.1830025.

Michel, Jean-Baptiste, Shen Yuan Kui, Presser Aiden, Aviva, Veres, Adrian, Gray, Matthew K, Pickett, Joseph P and Hoiberg, Dale (2011) 'Quantitative analysis of culture using millions of digitized books'. *Science,* 331(6014), pp. 176–182.

Rogers, Richard (2013) 'The website as archived object'. In *Digital methods*. Cambridge: MIT Press, pp. 61–82.

Samar, Thaer, Huurdeman, Hugo, Ben-David, Anat, Kamps, Jaap and de Vries, Arjen (2014) 'Uncovering the unarchived web'. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, DOI: 10.1145/2600428.2609544.

Schneider, Steven and Foot, Kirsten (2004a) 'The web as an object of study'. *New Media & Society*, 6(1), pp. 114–122.

Schneider, Steven and Foot, Kirsten (2004b) 'Crisis communication and new media'. *Society Online: The Internet in Context,* 137–153.

Siegl, Erica and Foot, Kirsten (2004) 'Expression in the post–September 11th web sphere'. *Electronic Journal of Communication,* 14(1–2), http://www.cios.org/EJCPUBLIC/014/1/01414.html.

Soboroff, Ian (2002) 'Do TREC web collections look like the web? In *ACM SIGIR Forum,* 36(2), pp. 23–31.

Thomas, Arthur, Meyer, Eric, Dougherty, Meghan, van den Heuvel, Charles, Madsen, Christine and Wyatt, Sally (2010) 'Researcher engagement with web archives – challenges and opportunities for investment'. *Joint Information Systems Committee Report*, first published online 1 August 10, http://ssrn.com/abstract=1715000.

Vaughan, Liwen and Thelwall, Mike (2003) 'Scholarly use of the web: what are the key inducers of links to journal web sites?'. *Journal of the American Society for Information Science and Technology*, 54(1), pp. 29–38.

Weltevrede, Esther and Helmond, Anne (2012) 'Where do bloggers blog? Platform transitions within the historical Dutch blogosphere'. *First Monday,* 17(2), first published online 6 February 12, DOI: 10.5210/fm.v17i2.3775.

Xie, Zhou Cheng and Barnes, Stuart J (2008) Web site quality in the UK airline industry: a longitudinal examination. *Journal of Computer Information Systems*, 49(2), pp. 50–57.

Dr Anat Ben-David is a post-doctoral researcher with the Web Archive Retrieval Tools project (WebART), University of Amsterdam. She holds a PhD from the Science, Technology and Society programme, Bar-Ilan University (2012). Her research focuses on the politics of web spaces and on web historiography.

Hugo Huurdeman is a PhD researcher at the University of Amsterdam, involved in the WebART project. His research focuses on information access, and explores retrieval tools and interfaces for web archives that facilitate analytical tasks. Hugo has an MSc degree in Information Science (2007) and an MA in Digital Library Learning (2012).