MARIA-DORINA COSTEA

REPORT ON THE SCHOLARLY USE OF WEB ARCHIVES

DIG HUM LAB

NetLab

Maria-Dorina Costea
Report on the Scholarly Use of Web Archives

ISBN 978-87-93533-01-1

Published by NetLab
Helsingforsgade 14,
8200 Aarhus N.

www.netlab.dk

Contents

Abstract

Purpose: For many years, web archiving communities have dedicated much of their time and effort to developing the necessary archiving technology and procedures. With the better establishment of archiving infrastructure, their focus has now increasingly turned towards enhancing web archive use and usability. In support of these efforts, this paper's aim is to provide an empirical perspective of how researchers currently engage with web archives, their needs in connection to this source, as well as to identify some of the reasons for their non-use.

Methodology: Data was collected using a mixed method approach, which consisted of an online survey in two Danish universities, and a series of semi-structured interviews.

Findings: Findings suggest that currently there is limited awareness in the humanities and social sciences research communities of the possibilities of using web archives as a scholarly source. At the same time, the results describe researchers' various use methods, some of the challenges they are facing when using the archives, as well as their suggestions for potential improvements.

Research limitations: Limitations address sample size and sampling method. Further research is necessary in order to obtain more robust and representative results.

Originality/value: This study aims to contribute to the presently limited research on the use of web archives in general, and their scholarly use in particular.

Keywords: web archives, scholarly use, digital heritage, digital humanities

Paper type: Research paper

1 Introduction and background

In a short time span of almost three decades, the web has become an integral part of contemporary life, particularly for Western society. It has developed into an important platform where on a regular basis, millions of people create, disseminate, and consume cultural productions. At the same time, due to its mutable and fast-paced character, information published on the live web is especially vulnerable to loss and change. Acknowledging the ephemerality, as well as intrinsic value of this new type of cultural heritage, many national libraries and archiving institutions have taken action towards its safeguarding and perpetuation, by regularly collecting and storing vast amounts of web materials.

Often containing billions of web pages, these web archives can represent invaluable resources for scholars in the social sciences and humanities. Through the systematic analysis of their content, they can reveal patterns, trends and relationships in data throughout time, as well as uncover changes that have occurred on the live web. For scholars looking to study the evolution of cultural and social phenomena from different perspectives, they can represent veritable mines of information. Nonetheless, despite the variety and sheer quantity of data, as well as numerous methods for their investigation, web archives continue to be an underused source in research (Hockx-Yu 2014, Dougherty et al. 2010, Brügger & Schröder 2017, p. 1).

The limited use of web archives has been attributed to date to several technical, circumstantial and policy issues. First, likely due to their relative novelty, researchers are often unfamiliar with web archives, as well as the tools and methods to engage with them (Hockx-Yu 2014). While web archiving institutions have taken steps over time towards engaging with users in order to make them aware of the existence of this source, as well as to understand their needs, this remains a process that needs to be done on a continuous basis (Ibid.). Second, a lack of research infrastructure that meets scholarly requirements means that certain phases of the research process, such as data searching and selecting, can become arduous. For instance, a majority of web archives only offer URL search, which can present limitations if the scholar does not know the exact URL of the website they are searching for. If they are seeking to find multiple web pages, URLs can only be inserted and searched for one by one, which can be a time-consuming process (Brügger & Schröder 2010, p. 11). An alternative to this has been the introduction of full-text search, such as in the cases of the Portuguese Web Archive, PANDORA or Netarkivet. However, establishing the algorithms that would rank the thousands or even millions of possible search results in a hierarchy that is relevant to the user remains an ongoing challenge for web archivists (Ibid). Third, access to web archives is often restrictive, because of legal frameworks protecting copyright and sensitive or personal data, several web archives requiring either special access provisions or being accessible only on-site (Hockx-Yu 2014, Nielsen 2016 p. 31). The potential unreliability of archived web materials as a scholarly source, due to incompleteness or inaccuracy of some of the archived materials (which is not always easily detectable) has been an additional challenge yet to be overcome.

Efforts are currently being made to increase and improve the use and access to archived web materials. While in the past, the web archiving and information management communities have focused their resources on building the necessary technological infrastructure and establishing processes for collection development, in recent years focus has started shifting increasingly towards the users (Hockx-Yu 2014). The International Internet Preservation Consortium was created in 2003 by the Internet Archive and 11 national libraries (Australia, Canada, Denmark, Finland, France, Iceland, Italy, Norway, Sweden, the U.K., and the U.S.A.) with a view to establish web archiving standards as well as create training materials and open source tools for crawling, collecting, indexing, replaying and analysing web content[1]. Research infrastructure projects such as NetLab[2] and BUDDAH[3] have been established with the purpose of facilitating extensive and effective use of web

[1] https://netpreserve.org/web-archiving/tools-and-software/

[2] http://www.netlab.dk/

[3] https://buddah.projects.history.ac.uk/

archive materials, providing scholars with the necessary tools and knowledge to aid them in their research, all the while promoting cross-organizational collaborations and knowledge exchange. At a European level, the RESAW[4] research infrastructure project aims to achieve a "borderless flow of information" between national web archives, that would permit researchers to make use of their content unrestricted by national barriers.

More empirical data about the use and requirements of researchers from web archives is needed in order to support these efforts. This study aims to gain insight into this problem, by focusing on the following research questions:

- Who are the users and non-users of web archives?
- What are the reasons why some researchers do not use them?
- How and why are web archives used for research?
- What are the data and research infrastructure needs of researchers using archived web materials?
- To what extent does the existing research infrastructure meet user needs?
- In which ways could the functionalities of web archives be improved to meet these needs?

2 Literature review

In preparation for the online survey and interviews, a literature review was conducted, which laid the groundwork for the survey and interview questionnaire design. The papers brought up many common themes and relevant points for further investigation.

2.1 Use methods

A compendium of possible uses, complete with case examples, has been composed by the International Internet Preservation Consortium (IIPC) and made available on their website.[5] The cases discuss using web archives for link analysis, outreach and education (such as in the use of web archives in physical or online museum exhibits), accountability and visibility for web content that no longer exists, text mining (the analysis of textual patterns and trends in large corpuses), as well as their use for analysis of technology trends. The various methods in which web archives can be used for research is also explored through a series of 12 interdisciplinary use cases in *The Web as History: Using Web Archives to Understand the Past and the Present* (Brügger & Schroeder 2017), featuring examples of web historiography, as well as

[4] http://resaw.eu/

[5] https://netpreserve.org/web-archiving/case-studies/

using web archives to study the evolution of multiple political and cultural phenomena.

2.2 Web archive user studies

Although web archiving is presently performed in many different countries, oftentimes on a large scale, only few studies of web archive users have been undertaken so far, of which three aimed specifically to map out the needs, attitudes and behaviour of scholars.

One of the earliest user studies was conducted in preparation for creating the content selection criteria, the interface, and search options of the Web Archive of the Netherlands (Ras & Bussel 2007). Potential users of the archive were identified, and the usability of search and access tools, as well as user satisfaction with selected content, were evaluated. The study found that research is the main reason for web archive use, and that users prefer full-text over URL, similar to the way they use Google. The study also stressed the importance of having search guidelines present on the main page, where users can easily access them. Other important findings were the necessity for integrating a more hierarchical presentation of results for full-text search, and the request for having metadata and other descriptions about the page, website and the archiving.

In their user study on the Portuguese Web Archive, Costa & Silva (2010) identified several functionalities preferred or expected by web archive users. Similar to participants in the study of the National Web Archive of the Netherlands, respondents preferred full-text search over URL search. Participants also named some of the functionalities they would like to see included in the web archive's services, among which a search engine for images, one for videos, as well as another for old news. Other proposed functionalities were seeing the evolution of a page or site, which one participant suggested could be done by having a side-by-side comparison between multiple versions of a page, but also auto-completion of search box queries and a personal workspace area where a user can access and manage their search history.

In a report by the JISC, Dougherty et al. (2010) analysed the different ways researchers engage with web archives, and provided an overview of the common challenges they are facing. The report was based on a series of interviews with various stakeholders from the web archiving community. Some of the researcher needs the study identified include *stabilized web objects* (i.e. they are able to function as reliable scholarly objects of study), the ability to define what the stabilized archived object represents in reference to the live web, access to representations of fine-grained features of web objects, and the ability to enrich and annotate these web objects. Furthermore, the report identifies three types of access that researchers need in order to make value of archived collections: *administrative* (metadata providing information to help manage the archived resource, such as when and how it was archived), *descriptive* (metadata for purposes of discovery and identification), and *contextual* (this does not refer to placing the object in its original context, but rather the ability for the research object to be found by means of its relationship to other objects in research projects). The report also advances a series of recommendations for further consolidation of web archive research infrastructures.

Another study, conducted by the British Library in 2012, also analysed the use of web archives by scholars, this time from the perspective of users as well as non-users, in an attempt to understand the reasons behind limited scholarly use (Hockx-Yu 2014). The participants were asked to evaluate web archives in terms of perceived research value, and sought to identify the content and access mechanisms that would be required for effective scholarly use. The study found that those who valued the archive the most were scholars interested in web history, statistics and digital preservation research. In addition, researchers expressed their desire for more images, rich media and blogs to be included in the archive’s collection. In her paper detailing the study, Hockx-Yu also outlines scholarly requirements from web archives (also in relation to new, digitally-engaged methods of scholarship such as the digital humanities) and proposes a list of basic requirements for web archives.

Perceptions of web archives and their use by researchers were also explored by Sterling, Chevallier and Illien (2012) in their study of the use of the French National Library web archive. Fifteen interviews were carried out with three different groups: researchers, professionals and "average users”. When asked about their service and information needs, researchers mentioned several issues, among which accessibility and transparency in the documentation of the collection process (such as the selection criteria), describing the archive in ways that allow researchers to differentiate between the collected sites (such as sites that link to a URL and the sites to which it has links, usage statistics such as number of visitors, views and its ranking in Google results, but also popularity and reputation), and using cooperation and large communities in identifying important sites for focused collections, as opposed to small groups of experts.

3 Methodology

Two instruments were used to collect quantitative and qualitative data: (1) an online survey, distributed via email and internal faculty newsletters and (2) semi-structured interviews.

3.1 Survey

3.1.1 Selection of participants

A survey invitation was sent out to the relevant academic staff of Aarhus University and the University of Copenhagen (UoC). All members of academic staff from the departments of the Faculty of Arts (n=767), as well as from three key departments in the School of Business and Social Sciences - the Department of Law (n=73), and the Department of Political Science (n=212) - at Aarhus University received the invitation via email. The survey was also successfully distributed in the newsletter of the UoC Faculty of Humanities (n=615). The UoC Faculty of Social Sciences also published a text in Danish and English about the survey in their newsletter, inviting researchers to participate. However, this invitation produced no responses. In total, survey invitations were sent to a number of 1667 researchers, professors, and PhD students. A low general response rate was expected, due to the fact that the survey came from a third party. A lower response rate was also expected from members of faculties that were notified of the survey indirectly, via newsletter.

3.1.2 Survey design

In addition to the survey link, the email invitation contained a brief description of the survey’s goals, as well as information about NetLab’s organizational mission. The survey was implemented using the Google Forms framework (forms.google.com). The questionnaire was kept short, in order to reduce abandon rate and increase the time users spent on questions (Chudoba 2011). The survey opened with a short description of its aims, and an introductory question asking participants if they have ever used web archives. Based on a yes or no response, respondents were then taken to one of two sets of questions.

Both user and non-user questionnaires contained multiple choice, linear grading scale and open-ended questions that resulted in a mixture of qualitative and quantitative data. By choosing this method, it was possible to gain a broad overview and employ standard analytics while at the same time also gaining a more nuanced perspective.

The user questionnaire aimed to gain insight into the methods used by researchers, the research questions and topics they use the archives for, as well as their perceived value of web archives in terms of usefulness to research. The users were also asked to rate a set of search, selection, visualisation, and extraction functionalities in terms of their importance, and offer their suggestions for the improvement of web archives. Non-users were asked to detail the reasons for not using web archives, and share their opinion about the value of web archives for research.

Before being sent out to participants, the questionnaire was tested with 3 different volunteers in order to ensure that all questions were fully understood.

The study used descriptive statistics to analyze the findings of the questionnaire.

3.1.3 Limitations

Respondents were able to decide for themselves whether to participate or not in the survey. For this reason, a degree of self-selection bias is likely to be present in the number of users who completed the survey compared to non-users, due to users' pre-existent interest in the subject. In order to minimize this effect, the invitation emphasized that responses from non-users were also important to the survey's aims.

3.2 Interviews

Semi-structured interviews were carried out, starting from a number of guiding questions. The aims of the interviews were to triangulate data, as well as potentially uncover issues that might have been overlooked in the online survey. Interview duration varied from 25 to 60 minutes, depending on the participants' interest and time availability. Selection of participants was done using purposive sampling. The participants included 5 researchers (3 web archive users and 2 non-users) of various ages, academic qualifications and research interests.

To analyse the semi-structured interviews, a thematic analysis method was used. The information was first coded, after which several themes were derived from the codes and reviewed.

3.2.1 Users

Users were asked to describe one or more research projects for which they have used web archives as one of their sources. The subsequent questions concerned 5 different topics: *data needs; search process*; *selection process*, *extraction* and *analysis*. The questions aimed to document the method that the researcher used during each phase of their research process, the challenges or limitations they faced in working with their different web archive sources, what workarounds they employed to overcome those challenges, as well as any suggestions for further improvement of the current research infrastructure.

3.2.2 Non-users

Non-users were asked to perform a first-time test search on the Internet Archive, using a research question of interest to them as a starting point. Both keyword and URL search were attempted in order to find relevant data, and the various functionalities of the Internet Archive were evaluated in terms of usability. Participants were also asked to suggest potential improvements that would provide them with a better user experience.

3.2.3 Limitations

Due to a small pool of potential participants, but also due to time and resource constraints, the end number of interview participants was limited. Further research is necessary in order to obtain more robust and complete results.

4 Results

4.1 Survey

In total, a number of 88 respondents filled in the survey completely or partially. This number was considered sufficient for the purpose of this survey, which sought insight, rather than generalization of results. Of the total number of participants, 40.9% (n=36) identified as users and 59.1% (n=52) as non-users.

Terminology

When asked about which web archive they use, 8 participants who identified as users named collections of digital and digitized content such as *LARM* (archive of Danish radio and television programme from the State Media Archive), *Mediestream* (Danish digital archive with audio-visual content and newspapers), *Lantmateriet.se* (the Swedish mapping, cadastral and land registration authority), but also *Instagram*, as web archives[6]. This suggests that the term “web archive” is for the moment not sufficiently self-explanatory, for a significant number of researchers, “web archives” representing an umbrella-term for archives found on the web, rather than simply collections of archived websites. This could be due to the relative recency of web archives, suggesting an ongoing lack of audience familiarity with the source.

Due to the fact that this survey is particularly aimed at users of collections of archived websites and webpages, the eight responses were not factored into the final survey results. The total number of web archive user responses was 28 (31.8%), 8 (9.1%) respondents were users of other digital collections, 52 (59.1%) were non-users.

Demographic characteristics

Many web archive users listed *History* (18.9%, n=7) as one of their research areas, followed by *Archaeology* (13.5%, n=5) and *Linguistics and Languages (10.8%, n=4)*. For non-users, Linguistics and Languages was cited most often as a research field (18.9%, n=16), followed by Anthropology (13.1%, n=8), Education (9.8%, n=6) and Politics (8.2%, n=5). A complete overview of the research fields is found in Figures 1 and 2. Figure 3 describes the different age groups of participants in the study.

[6] A note defining web archives was later introduced in the survey description.

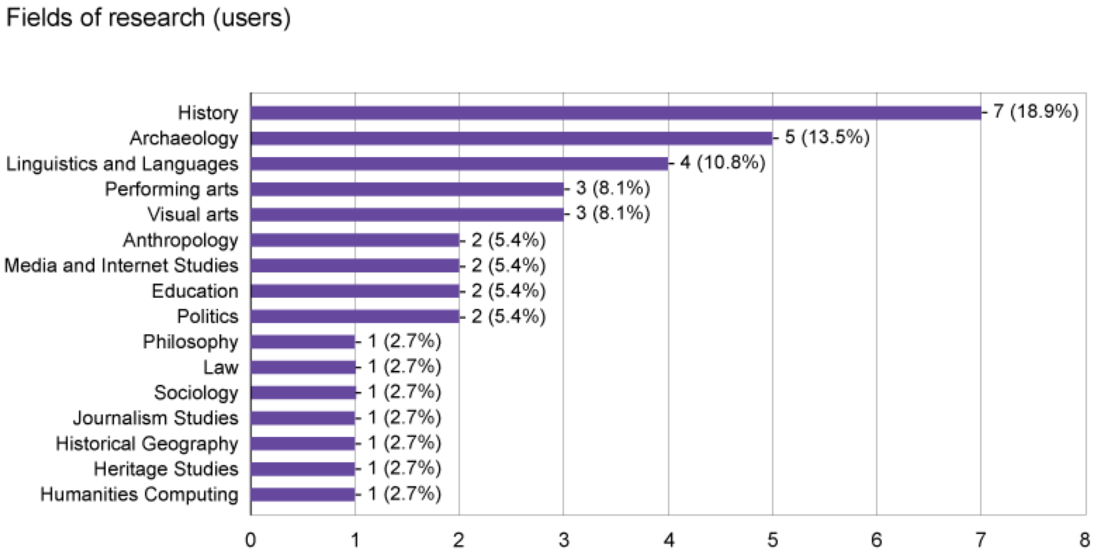
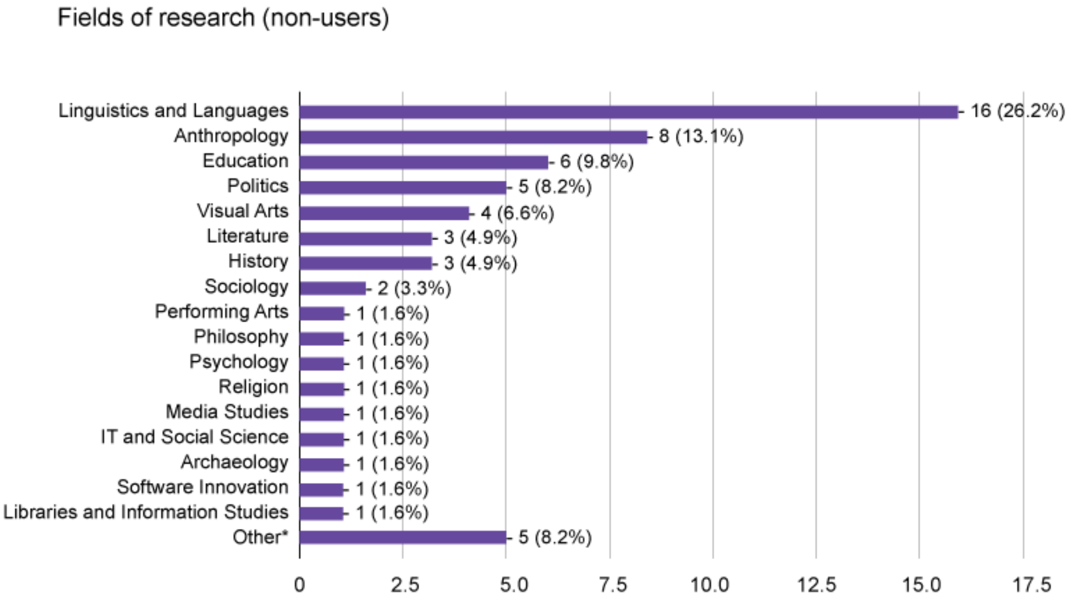


Figure 1. Fields of research (users)



*Tourism (1), Architecture (1), Cultural Studies (1), Organizational Practice (1), Environmental Policy (1).

Figure 2. Fields of research (non-users)

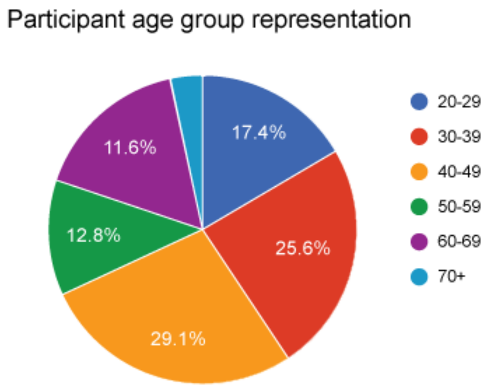


Figure 3. Participant age group representation

Awareness of Netarkivet as a potential source for research (users and non-users)
A relatively new web archive (established in 2005), Netarkivet, the National Danish Web Archive, has been accessed so far by 123 unique online users (researchers with a Ph.D degree) and 15 inhouse users (Schostag 2018, personal communication, 9 Feb). Recently, Netarkivet has revised their interpretation on the law on who can receive access, including now also students writing their Master thesis. The study aimed to find out if researchers were aware that Netarkivet is available to researchers as a potential source. A majority of the participants in the survey (of 88 responses, 59%, n=52) said they did not know that copies of Danish websites are achieved by Netarkivet and made accessible to researchers. Of the researchers already using web archives, half (14 of 28 responses) knew about the existence of Netarkivet.

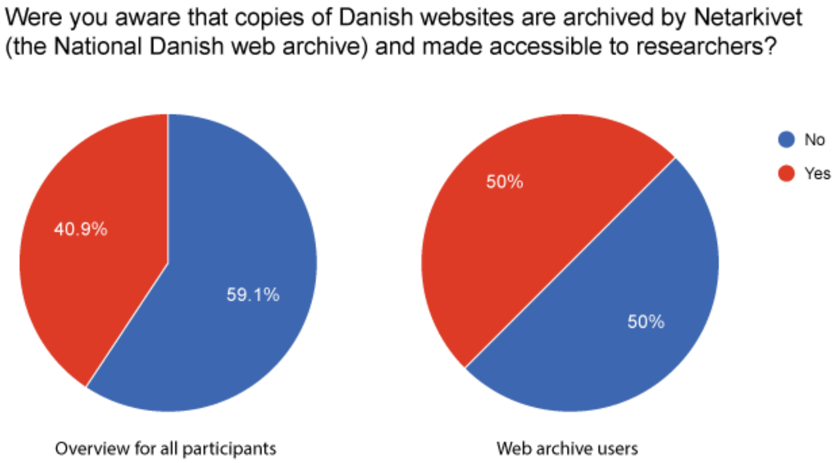


Figure 4. Awareness of Netarkivet as a potential source

4.1.1 Non-Users

Reasons for not using web archives in research

A majority of respondents (32.7%, n=17) answered that the reason they don’t use web archives in their research is that they do not find the content relevant. However, a combined number of 25 participants (48.1%) attributed their non-use to issues related to lack of information. This included not knowing that web archives existed (25%, n=13) or not knowing how to use them or what they contain (23.1%, n=12). A smaller number (7.7%, n=4) said they knew about the existence of web archives, but they hadn’t come to their minds as a potential source for research. Three respondents (5.8%) answered they do not use the source because they do not think it can provide them with sufficient or accurate information. Other cited reasons for non-use were “copyright and privacy law difficulties with offline use”, and not having had the opportunity due to just starting research. One respondent answered that they haven’t had the opportunity to use web archives in their research, but had used them to find materials for teaching.

Perceived value of web archives

While a majority of respondents said they do not know enough about web archives to give a response (39.2% n=20), those who did, generally found value in them for a number of different reasons. Eleven of the respondents (21.6%) evaluated them positively (4 or 5 on a 1-5 scale), due to their role as repositories of digital heritage: “much material in the field are nowadays digital, and it disappears or change [sic] address, so much documentation is lost”; “It is a very central medium in our contemporary culture”; “historical data on how arguments and trends unfolded are precious and valid data for research”; “If the right material is archeived [sic], also current material, it would be very helpful”; “It may be important to be able to access the knowledge that is no longer available, and some projects are dependent on being able to conduct the kind of archeology [sic] in our digital culture. But only 4 and not five because it is not all research that is dependent on online archives. There are actually research fields that do not need that resource.”; “if the material is relevant, it is just a new platform for the relevance”. One respondent also thought that web archives would be able to help him easily go through large amounts of data, as well as easily retrieve the information he seeks. An additional 2 respondents (3.9%) offered a positive evaluation, but did not comment on their choice.

A more neutral outlook (3 on the 1-5 scale) came from respondents that appreciated the richness of data, but were also aware of potential issues and challenges they might be faced with, such as difficulty of use or access, as well as potential gaps in data (13.7%, n=7): “Existence is a plus, difficulty to use a drawback”; “Overall, I think they are useful, but it takes time to access it, and a lot of visual material is not available - to my knowledge.” One concern was also a potential lack of data representativeness due to the way data is selected: “It is the archive administrators that decide what to store in the archive. The content maybe not be representative or could be misleading somehow.” An additional 5 respondents gave the same evaluation (9.8%), but did not provide comments.

A low score (1 or 2) was given by participants (5.9% n=3) who evaluated the value of web archives in relation to personal relevance, stating that their research is field-based. Another 3 participants (5.9%) gave the same evaluation, but did not provide comments.

4.1.2 Users

Most frequently used web archives

The Internet Archive Wayback Machine (%41.9, n=18), was the most often cited web archive used by researchers. This was followed by Netarkivet (11.6%, n=5), the US Library of Congress Web Archive (11.6%, n=5), and the US Web Archive (7%, n=3). 8 respondents (18.6%) provided the names of other types of digital collections.

Web archive use patterns

A significant majority of the users responded that they currently use or have used web archives with qualitative methods of research (e.g. analysing individual website content) (75.9% n=22). 3 users (10.3%) had used web archives for quantitative studies, while 4 (13.8%) had used both. Other uses that were mentioned were finding material for teaching (curriculum development), obtaining basic information on a certain subject, or for various experiments. The predominant methods for using web archives were history, and textual analysis such as political discourse analysis and multimodal discourse analysis. Research topics included:

- Music history
- Organizational and personnel change in specific public offices
- Development of specific websites
- Development of style conventions
- Political discourse analysis
- Memory studies
- History of Danish radio broadcasting (DR web pages)
- How visual coverage of politics changes over time
- Church archaeology
- Social work research: finding websites of earlier municipal institutions
- Historical research
- Researching public comments pages
- Research on creative policy materials
- History of administration/legislation
- Children’s media history – web pages for children
- Communication of design on social media

Perceived value of web archives

Similar to non-users, a majority of scholars who use web archives rated them positively (4 or 5 on a 1-5 scale) in terms of usefulness to research (56% n=14). When asked to provide the reasons for their evaluation, participants who had rated them highly stated they did so due to the importance of web archives as a source for studying contemporary society: “Web archives are essential if working on an issue related to what has been going on in the world for the last 20+ years…”.; “The internet is emerging as a general archive of all text”. Web archives were also considered highly useful due to their ability to provide information that is not on the web anymore: “There are lots and lots of websites that disappeared from the current web containing important and unpublished information about my research field.”, but also for visualising the evolution of phenomena.

Participants who chose a midpoint or lower score (3-2 on a 1-5 scale) (44% n=11) argued that while the content of web archives is very valuable to their research, there are a number of different issues which currently detract from their usefulness. The different issues and challenges brought up by the users can be classified into three themes, concerning (1) data reliability, (2) access, and (3) usability capabilities of web archives.

1. Data reliability, completeness and integrity

An issue that was mentioned several times was the incompleteness or limited reliability of data (16% n=4). One participant who had used Netarkivet mentioned a lack of music, while another mentioned that pictures they searched for on the Internet Archive were not stored in sufficiently high resolution for them to use. One participant using the Internet Archive and PANDORA mentioned the need for more transparency regarding the algorithms behind the archiving and search processes: “An interface online is not enough and search results need to be reliable. Explain the algorithms behind archiving and search to researchers.”

2. Access

One user of Netarkivet mentioned difficulties in data access and mobility: “The potential usefulness of the data is 5 (Extremely useful), but data access and mobility is 1 (Not useful). The sheer amount of data means that access and mobility is extremely important, because, as a researcher, I need to be able to sample and compute on a HPC infrastructure that fits my needs. In my case, I typically need GPU accelerated computing at a scale that are only offered by very HPC centers in DK.” For this, the user suggested a possible resolution through the creation of database dumps for specified time slices that can be accessed at research facilities in Denmark.

3. Usability

Several users found web archives to be presently only moderately useful or below, due to an only limited usability of their functionalities (20% n=5). In addition to general comments on the usability of facilities, participants also mentioned specific issues. One user of the Internet Archive mentioned a difficult search process, while another mentioned a lack of structured access options. One user of the Internet Archive and PANDORA mentioned a need for export facilities and common file standards to work with data.

Evaluation of potential functionalities

1. Search functionalities

Of the suggested search functionalities, the inclusion of a *search engine for images* in the web archive was considered the most important, receiving a weighted average of 3.92. An average above 3 (between *important* and *very important*) were also given to the ability to *search for more URLs in one search* (3.31) and a *search engine for videos* (3.06). *User search history* and *auto-completion of search fields* received an average score of 2.29 and 2.81 respectively.

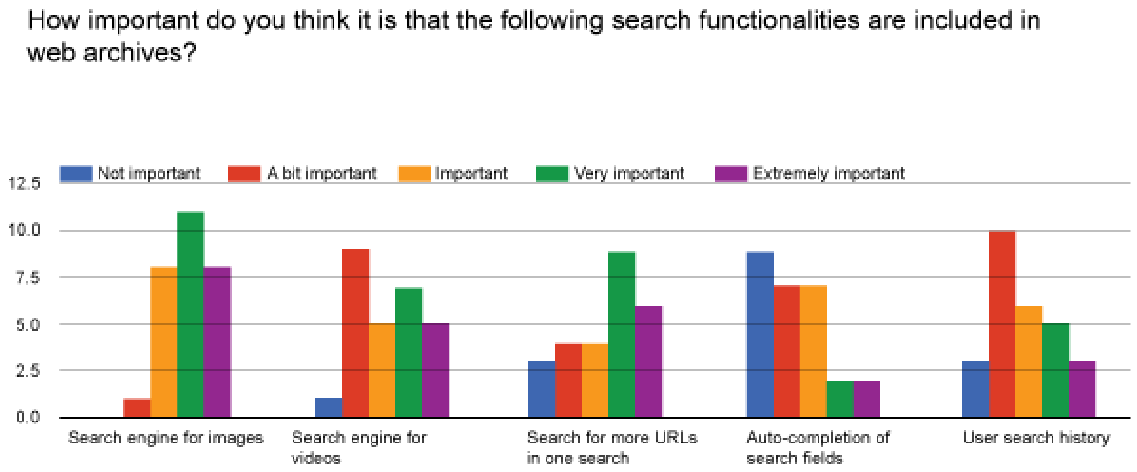


Figure 5. User preferences for potential search functionalities[7]

2. Data selection, visualisation and extraction functionalities

The functionality considered most important by users was *the ability to create and extract a corpus from a larger collection*, with a weighted average of 4.03. All other proposed selection, visualisation and extraction functionalities were rated above 3 in terms of importance: *the possibility to download data in multiple formats* (3.88); *tools that help with the discovery of patterns, trends or relationships in data* (3.61); *side-by-side visualization of a website throughout the years (the ability to compare different versions of a website)* (3.51); *collections of archived websites specific to my research area* (3.22).

[7] Expressed as number of individuals.

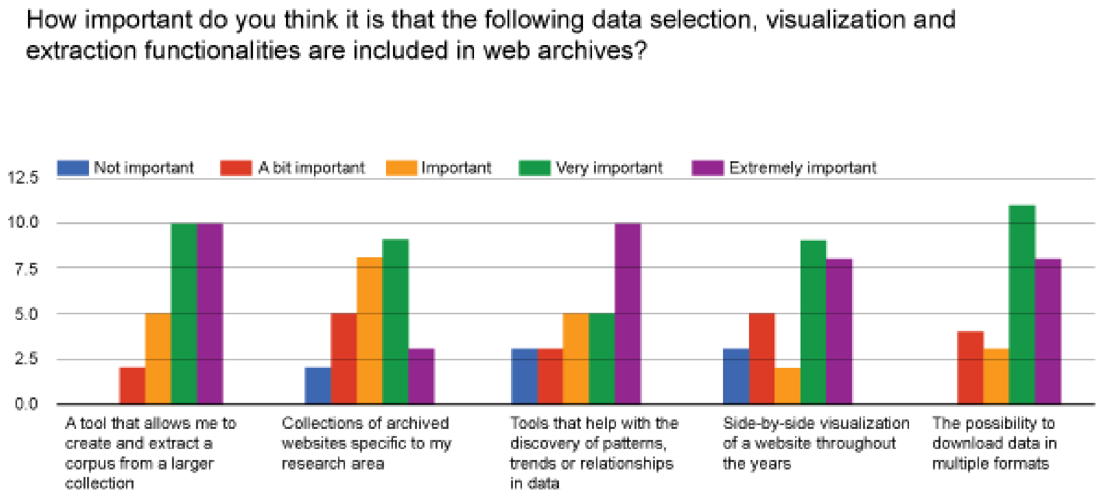


Figure 6. User preferences for potential data selection, visualisation, and extraction functionalities[8]

[8] Expressed as number of individuals.

4.2 Interviews

4.2.1 User interviews

	Academic profession / Degree	Research Interests	Web archives used	Short description of project
P1	Assistant professor	Media history, web historiography	Netarkivet; Internet Archive	P1 has used web archives in a cross-media analysis of the historical evolution of the Danish Broadcasting Corporation's role as an educator, by conducting close readings of multiple cases illustrating how educational programs have been represented on DR's website. Using both analogue sources as a point of reference and searches within Netarkivet and the Internet Archive (for materials older than 2005), the participant was able to successfully identify several case studies relevant to the research. Nevertheless, the participant has described the process as very challenging, due to a series of issues relating to discoverability and incompleteness of data, as well as absence of metadata and documentation.
P2	Ph.D. fellow	Media policy, program schedule management, media sociology	Netarkivet	P2 has used Netarkivet in an analysis of the development that takes place when public service broadcasters have to switch from regular broadcasting (radio and television) to digital media, and interactions that occur between different platforms. For the project, P2 needed access to the online television service of DR from 2010, in order to describe the content of the page (for example, the presentation of the different channels). The participant used this source in conjunction with screenshots of the live site for later periods, as well as analogue archive materials. The chosen approach was to search for dr.dk/tvnu and choose one week for 2010. An analysis was carried out for the single capture from that week. The participant described the experience of using the archive as overall satisfying, but not with limitations related to the availability of dynamic content, such as in the case of the children's TV channel, but also finding out which page was most appropriate to use.
P3	Research librarian	Music history, music and technology	Netarkivet; Internet Archive	P3 has used web archives in numerous music history projects, such as researching the so-called "digital music revolution", for the period 2005-2012. The participant has described using web archives as one of many other sources (for example, Mediestream or Youtube), for building data on particular topics. In the approach to search for relevant materials and find new, interesting data, the participant has used both keywords and URL search, as starting points. The user has described his research as a non-linear process, as "working in explosions", and while using Excel in order to keep track of the many selected results has proven very helpful in absence of other options, the user believes an interface where results could be saved for later use and management would prove more effective.

Table 1. Profiles of participants

Documentation, data and metadata needs

In the first stage of the user interviews, participants were asked to provide a broad account of projects where they used web archives as a primary source, with a focus on describing their particular data, metadata, and documentation needs, and the extent to which they were met by the web archives. In regard to data needs, the participants, who had used methods such as history and discourse analysis, had required integral web pages, in order to carry out a full and correct reading of the different discursive components, an aspect which was not always possible. In addition, one participant pointed out a near absence of audio-visual materials, as well as poor quality of social media content, as a particularly challenging aspect for his research purposes (P3). The lack of accessibility to a part of the video content that has indeed been archived has been seen as an additional problem: “I know that since 2012 videos have been collected from YouTube [by Netarkivet], but you can’t search for them in the interface right now” (P3).

While the participants did not expect the data to be always complete, they did, however, voice their need for information that would provide an accurate picture of their research object (e.g. which [if any] parts were missing): “Unless we can specifically see that the picture is missing or something, it's very difficult to know whether something's not there and how the harvesting settings might have influenced what it is that we're studying.” (P1). Metadata and documentation of any factors that could influence the content, aspect and behaviour of archived material were mentioned as necessities that could improve the scientific reliability of the objects (P1, P3). Nonetheless, due to the perceived value and indispensability of the archived material, researchers were determined to find creative ways of working around existing metadata accessibility issues. One such approach was analysing the HTML code of archived web pages, which provided the user with metadata and original page names that are otherwise untraceable: “within the HTML code, especially for 2011-2012, you could actually extract the original internet addresses from that short address[9] […] Studying the code gives you so much more. It helps you if you have at least some technical skills” (P3).

The search process

All participants used a combination of sources in order to gain a fuller picture of the investigated cases, either combining digital and analogue sources, or using multiple web archives (two of the participants had used Netarkivet in combination with the Internet Archive), to cross-examine materials or to complement missing data (either websites or web pages in their entirety, or web objects from an existing page). Participants also mentioned using additional analogue sources as reference points in searching for specific materials within the archive (P1, P2,

[9] Referring to redirect URLs such as bit.ly or Tiny URLs

P3). This was largely attributed to the absence of an overview of materials existing in the web archive and how they might differ from each other. As one participant explains: “I would get thousands of results. So in that way it’s not really possible to find something representative. That would be huge work. I went from the radio and television [archive] and found some programs there that I wanted to look at, and then I saw how they were represented on the website.” (P1). One participant (P3) mentioned preferring to search the Internet Archive prior to any use of Netarkivet, due to its ease of access.

The selection process

While using web archives for a limited amount of time and a narrow purpose has been described as little problematic (such as in the case of P2, who only used one capture of a frequently and extensively archived page), users have described extensive use of archived materials for close readings as quickly becoming an onerous process. This issue has been attributed by participants to the current legal barriers preventing them to extract data, combined with the absence of tools that would permit an effective management of the high volume of information. In dealing with this issue, several researchers mentioned they have improvised workarounds of constructing collections and content management systems outside the archive (such as folders of screen dumps or Excel files that keep track of the materials), but at the same time made known their hope for a future “workspace” within the web archives that would permit them to easily and quickly collect and analyse the material. (P1, P3). Selecting material that is meaningful and comprehensive from thousands of search results, with little reference of what underlying differences might be between similar results, is another obstacle that has been brought up by users (P1, P2, and P3). In this case, there are currently few criteria in the archives and little related documentation that can help researchers make an informed decision.

Extraction and analysis

Legal frameworks, such as the Copyright Act and the Act on Processing of Personal Data, currently regulate the extent and way the content of web archives in Denmark is used. Content from Netarkivet may not be reproduced, with a few exceptions such as personal use, educational purposes, or scientific presentations and publications[10]. According to Netarkivet’s terms of use, creating copies of the database in digital form is currently not permitted (Netarkivet, 2018). In this context, researchers undertaking page-by-page analysis have generally resorted to compounding collections of screenshots for study, using software tools such as *Paparazzi*[11], which enable them to capture

[10] A full description of Netarkivet’s terms of use is available at http://netarkivet.dk/adgang/

[11] https://www.macupdate.com/app/mac/15966/paparazzi

also parts of the website that do not fit the screen, and would otherwise require scrolling and multiple snapshots.

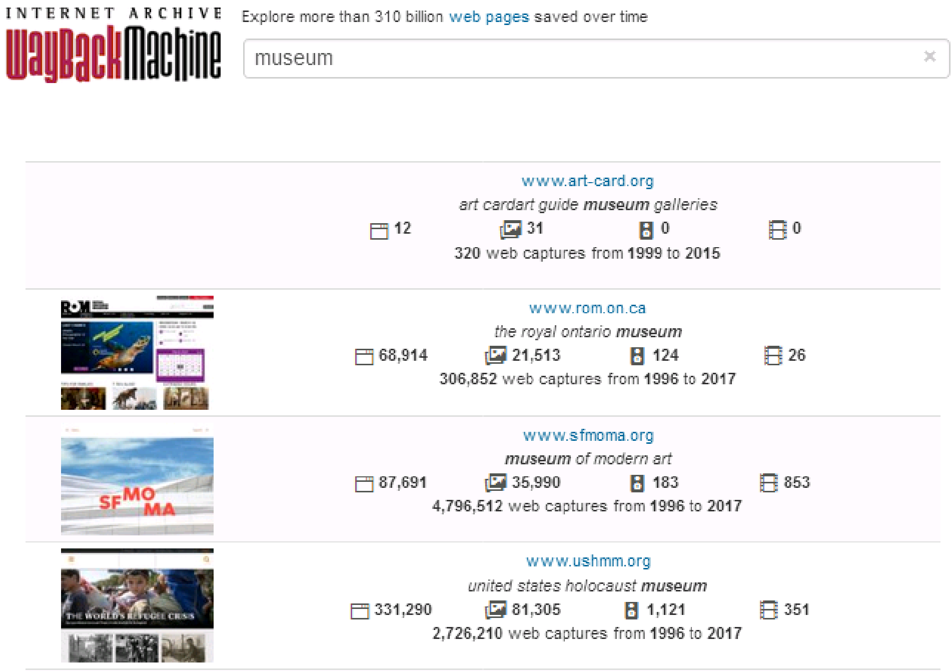
When asked about whether they had tried to carry out quantitative analysis involving web archive corpora, only one participant had done so (P1). However, the project had been put on hold due to technical issues. In the case of Netarkivet, extracting datasets from the archive for analysis is also currently not possible, due to legal protections on the archive's content. Nonetheless, large-scale, quantitative analyses can be still performed within the premises of the Danish Royal Library, using the DeIC National Cultural Heritage Cluster, an e-science infrastructure that enables high-performance computing of vast quantities of data.

4.2.2 Test with first-time users

	Academic profession / Degree	Research Interests	Short description of project
P4	M.A. student, 4th semester (Thesis research)	Heritage management, production and communication of heritage	Critical discourse analysis on materials related to the cast collection of the Statens Museum for Kunst (SMK). The participant wishes to study the evolution of the way SMK has portrayed their cast collection in the online media, as part as a wider study.
P5	Ph.D. Fellow	Historical Anthropology, Cultural Anthropology and History of Technology	The participant did not have an ongoing research project in which web archives were included as a source. The test was carried out based on the participant's previous research interests (religious pluralism in the Arab world, and urban development and planning in the Arab world before and after colonialism).

Table 2. Profile of test participants

For the first task, participants were asked to perform a free text search, using keywords relevant to their research topic. P4 chose to search for “museum”, while P5 searched for news websites from Egypt, by using the keywords “journalism” and “Egypt”. A first comment and suggested point of improvement concerned access to the different types of archived objects. Each search result contained a summary including the number of distinct web pages, image files, audio files and video files, in addition to the number of web captures for the result and time period for which the captures had been harvested. However, the different categories only comprised the number of items, and did not further link to overviews of the actual items, which is what the participant had initially expected and suggested would be highly useful to have available in the future.



INTERNET ARCHIVE WaybackMachine

Explore more than 310 billion web pages saved over time

museum

www.art-card.org
art cardart guide ***museum*** *galleries*
12 31 0 0
320 web captures from 1999 to 2015

www.rom.on.ca
the royal ontario ***museum***
68,914 21,513 124 26
306,852 web captures from 1996 to 2017

www.sfmoma.org
museum *of modern art*
87,691 35,990 183 853
4,796,512 web captures from 1996 to 2017

www.ushmm.org
united states holocaust ***museum***
331,290 81,305 1,121 351
2,726,210 web captures from 1996 to 2017

Figure 7. Summaries of archived content for "museum" search results

An additional issue was a lack of options to filter results. In the case of P5, additional time was spent manually selecting results according to their top-level domain, in order to find the ones that corresponded to the researcher's needs (in this specific case, news websites from Egypt).

Several inconsistencies were noticed in the summary data. First, the summary stated that 0 distinct videos had been archived for the respective domain. This was a confusing aspect for the noticing participant, since the respective website did contain several working videos. Second, the summary for one of the results, (www.statensmuseumforkunst.dk) displayed that it had been captured between 2000-2007. However, when opening the URL in the calendar view, it was described as having been saved 57 times between November 9, 2000 and June 29, 2017, showing a discrepancy of 10 years between summary statements.

During the second part of the test, users searched for a website of their choice using URL search, and then explored the different sections of the website overview. P5 suggested that more clarity is necessary concerning the search process, stating that she had not noticed one has to enter a full URL in order to search for a specific website. More coherence and clarity were additionally requested in terms of how to use the archive in general. P4 remarked that use information was spread out in small notes on the site (such as the note concerning that the calendar view maps number of times the site was crawled, not how many times it was updated), but was unable to find a comprehensive guide on how to use the web archive (Information in the FAQ section was not systematic, and contained combined information about the Internet Archive in its entirety[12], making it difficult to pinpoint the questions and answers describing exclusively how to use the web archive).

While both participants found the calendar view helpful and easy to navigate, one suggestion was given in connection to the way the calendar is displayed. P4 proposed a secondary viewing possibility, which would display thumbnails of the websites for each date, in order to quickly spot any major changes that have occurred throughout time. For P4, this functionality was considered highly relevant in pinpointing key moments in the National Gallery's discursive evolution concerning their cast collection, which the participant identified as her main research focus. For the URL search task, P5 attempted to find city maps, images and historical documents within the British National Archives website. While the participant stated she appreciated "an extra layer" of information (referring to the many temporal dimensions of the archive), she quickly found that other sources were more appropriate and accessible in this case.

For the "Summary of domain" section in the website overview, P4 had expected to find a complete overview of the archived content of the domain, similar or related to the keyword search summary (see above). However, the content of the domain consisted of other types of information (e.g. "Summary on MIME-types Count" or "Key Summary for the TLD/Host/Domain"), which was considered highly technical ("For me it is extremely bare. You would probably need to be a web developer to understand this.") and was appreciated by the participant as not usable or relevant for most casual users. The participant also suggested that it would be useful to have a hover function explaining the different jargon terms, such as "capture", which could also potentially be enabled or disabled, according to the needs of each user.

Finally, participants explored the playback views of the website. P4 believed comparing the different temporal versions of the website would be very useful to their research. When asked about how they would proceed to do this, the participant stated that she would do screenshots, in order to be able to compare the content and see when and what changes had occurred. A further suggestion came regarding missing content on the website. When browsing the page, one of the videos was not working for a particular date, having a screenshot in place. It was later revealed that the video was in fact working for future dates. The participant suggested it would be very helpful to have a "bubble" stating the last date the video had been archived, if at all, but also a possible mention in the same form of why videos, but also other content is not working or missing.

[12] The Internet Archive comprises to date multiple categories, including web, texts, video, audio, software and image archives.

5 Conclusion and perspectives

This study set out to identify current ways scholars in the humanities and social sciences engage with web archives, and to describe the key aspects that they believe would facilitate and increase their use. At the same time, it aimed to pinpoint some of the current non-user perceptions of web archives. The study sought to be explorative, and thus the results are not fully conclusive or representative for the population as a whole.

Results suggest that a significant segment of the research community in the humanities and social sciences is still unaware of the existence of web archives, but also many do not know exactly what they contain, or how they can be used as a source for research. In this sense, informing researchers about the opportunities of web archives could increase their use, but could also contribute to the growth and development of the communities supporting web archive research activities.

While in general both users and non-users have appreciated the value of archived web content highly, there are several key areas of improvement that researchers strongly require, in terms of access to metadata and documentation, better discoverability options for the archived content, data selection and management, as well as better access to more ways of analysing the data.

In the case of using the materials for close readings, many technical, as well as legal limitations currently underpin the efficiency of the research process and reliability of results. Not knowing which objects might be missing from the page, or how harvesting settings might have influenced the way results are displayed has been a significant issue for the scholarly method of users. Comprehensive documentation and metadata is needed in order to ensure the scholarly validity and trustworthiness of archived objects, and to provide a picture of research objects that is as complete as possible. Being able to search across multiple archives at the same time, similar to the way the Memento[13] project currently operates with some archives, has also been seen by several participants in the study as a way forward to improving the issue of data incompleteness.

The possibility to search for different file types both within archives and within a selected domain would greatly help researchers seeking a specific type of content. This issue has been particularly pressing for researchers interested in certain types of data, such as audio-visual files in the case of music historians. Furthermore, researchers are currently encountering great difficulties while trying to discriminate between the vast amount of search results, and to choose the most appropriate

[13] Memento is a project funded by the United States National Digital Information Infrastructure and Preservation Program (NDIIPP) that aims to improve the discoverability of archived web content. It permits users to transition between web archives in order to find the best version of a page for the time they are looking for. More information at: http://mementoweb.org/about/.

option to go with. The possibility to compare and contrast the characteristics of different results, as well as having visual overviews of each capture in order to quickly spot changes in webpage content throughout time, have been suggested by participants as possible solutions to this issue.

Finally, highly important both to participants in the survey and interviews has been researchers' ability to define corpora for their own specific research needs. Participants have expressed this need both in terms of extracting large datasets to perform data analytics, and for hand-selected collections. In the latter case, a platform integrated in the archive could potentially serve both as a data management system and a tool for comparison between selected items.

Bibliography

Brügger, N. & Schroeder, R. (eds.) 2017. *The Web as History: Using Web Archives to Understand the Past and the Present*. UCL Press, London.

BUDDAH (Big UK Domain Data for the Arts and Humanities), 2017. *BUDDAH Home Page*. Available from: https://buddah.projects.history.ac.uk. [11 November 2017].

Costa, M. & Mário J. S., 2010. "Understanding the information needs of web archive users." *Proc. of the 10th International Web Archiving Workshop*. Vol. 9. No. 16.

Chudoba, B. 2018. *How much time are respondents willing to spend on your survey?* Available from: https://www.surveymonkey.com/curiosity/survey_completion_times/ [9 February 2018].

Dougherty, M., Meyer, E.T., Madsen, C.M., Van den Heuvel, C., Thomas, A. and Wyatt, S., 2010. "Researcher engagement with web archives: State of the art". Joint Information Systems Committee Report, August 2010. Available from SSRN: https://ssrn.com/abstract=1714997 [14 February 2018].

Hockx-Yu, H., 2014. "Access and scholarly use of web archives". *Alexandria*, 25(1-2), pp.113-127.

IIPC, 2015. "IIPC GA 2014 Paris - Researchers and Web Archives". YouTube video, 18th June. Available from: https://www.youtube.com/watch?v=nqbr7tDIMUk [6 February 2018].

IIPC, 2017. *Case studies*. Available from: https://netpreserve.org/web-archiving/case-studies/ [11 November 2017].

IIPC, 2017. T*ools and software*. Available from: https://netpreserve.org/web-archiving/tools-and-software/ [11 November 2017].

Nielsen, J., 2016. *Using Web Archives in Research – An Introduction*. NetLab. Available from: http://www.netlab.dk/wp-content/uploads/2016/10/Nielsen_Using_Web_Archives_in_Research.pdf [14 February 2018].

NetLab, 2017. *NetLab Home Page*. Available from: http://www.netlab.dk/ [11 November 2017].

Memento Project, 2018. *About the Memento Project*. Available from: http://mementoweb.org/about/ [14 February 2018].

Ras, M. and van Bussel, S., 2007. *Web archiving user survey*. Technical report, National Library of the Netherlands (Koninklijke Bibliotheek). Available from: https://www.kb.nl/sites/default/files/docs/kb_usersurvey_webarchive_en.pdf [14 February 2018].

RESAW (A Research Infrastructure for the Study of Archived Web Materials), 2017. *RESAW Home Page*. Available from: www.resaw.eu. [09 February 2018].

Stirling, P., Chevallier, P. and Illien, G., 2012. "Web archives for researchers: Representations, expectations and potential uses". *D-Lib*, *18*(3/4). Available from: http://www.dlib.org/dlib/march12/stirling/03stirling.html [14 February 2018].