# WARCnet

## Special Reports

# Skills, Tools, and Knowledge Ecologies in Web Archive Research

# Skills, Tools, and Knowledge Ecologies in Web Archive Research

*Sharon Healy (Maynooth University), Helena Byrne (British Library), Katharina Schmid (Bavarian State Library), Nicola Bingham (British Library), Olga Holownia (International Internet Preservation Consortium), Michael Kurzmeier (Maynooth University), Robert Jansma (University of Siegen)*

Sharon.Healy@mu.ie


WARCnet Special Report
Aarhus, Denmark 2022

Sharon Healy, Helena Byrne, Katharina Schmid, Nicola Bingham, Olga Holownia, Michael Kurzmeier, Robert Jansma: *Skills, Tools, and Knowledge Ecologies in Web Archive Research*


Published by the research network WARCnet, Aarhus, 2022.

Editors of WARCnet Special Reports: Niels Brügger, Jane Winters, Valérie Schafer, Kees Teszelszky, Peter Webster, Michael Kurzmeier.

Cover design: Julie Brøndum, Kamilla Rosenberg, Emma Lund Nielsen, Thea Laugesen
ISBN: 978-87-94108-09-6

WARCnet
Department of Media and Journlism Studies
School of Communication and Culture
Aarhus University
Helsingforsgade 14
8200 Aarhus N
Denmark
warcnet.eu

The WARCnet network is funded by the Independent Research Fund Denmark | Humanities (grant no 9055-00005B).

## WARCnet Papers

Niels Brügger: *Welcome to WARCnet* (May 2020)

Ian Milligan: *You shouldn't Need to be a Web Historian to Use Web Archives* (Aug 2020)

Valérie Schafer and Ben Els: *Exploring special web archive collections related to COVID-19: The case of the BnL* (Sep 2020)

Niels Brügger, Anders Klindt Myrvoll, Sabine Schostag, and Stephen Hunt: *Exploring special web archive collections related to COVID-19: The case Netarkivet* (Oct 2020)

Friedel Geeraert and Nicola Bingham: *Exploring special web archives collections related to COVID-19: The case of the UK Web Archive* (Nov 2020)

Friedel Geeraert and Barbara Signori: *Exploring special web archives collections related to COVID-19: The case of the Swiss National Library* (Nov 2020)

Matthew S. Weber: *Web Archives: A Critical Method for the Future of Digital Research* (Nov 2020)

Niels Brügger: *The WARCnet network: The first year* (Jan 2021)

Susan Aasman, Nicola Bingham, Niels Brügger, Karin de Wild, Sophie Gebeil and Valérie Schafer: Chicken and Egg: *Reporting from a Datathon Exploring Datasets of the COVID-19 Special Collections* (Dec 2021)

Emily Maemura: *Towards an Infrastructural Description of Archived Web Data* (May 2022)

Valérie Schafer, Jérôme Thièvre and Boris Blanckemane: *Exploring special web archives collections related to COVID-19: The case of INA* (Aug 2020)

Niels Brügger, Valérie Schafer, Jane Winters (Eds.): *Perspectives on web archive studies: Taking stock, new ideas, next steps* (Sep 2020)

Friedel Geeraert and Márton Németh: *Exploring special web archives collections related to COVID-19: The case of the National Széchényi Library in Hungary* (Oct 2020)

Friedel Geeraert and Nicola Bingham: *Exploring special web archives collections related to COVID-19: The case of the IIPC Collaborative collection* (Nov 2020)

Caroline Nyvang and Kristjana Mjöll Jónsdóttir Hjörvar: *Exploring special web archives collections related to COVID-19: The Case of the Icelandic web archive* (Nov 2020)

Sophie Gebeil, Valérie Schafer, David Benoist, Alexandre Faye and Pascal Tanesie: *Exploring special web archive collections related to COVID-19: The case of the French National Library (BnF)* (Dec 2020)

Karin de Wild, Ismini Kyritsis, Kees Teszelszky and Peter de Bode: *Exploring special web archive collections related to COVID-19: The Dutch Web archive (KB)* (Nov 2021)

Michael Kurzmeier, Joanna Finegan and Maria Ryan: *Exploring special web archives collections related to COVID-19: The National Library of Ireland* (Feb 2022)

## WARCnet Special Reports

Sharon Healy, Helena Byrne, Katharina Schmid, Nicola Bingham, Olga Holownia, Michael Kurzmeier, Robert Jansma: *Skills, Tools, and Knowledge Ecologies in Web Archive Research* (September 2022)

All WARCnet Papers and Special Reports can be downloaded for free from the project website warcnet.eu.

# Author Information

**Sharon Healy**, the lead researcher, is a final year PhD Candidate and GOIPG Irish Research Council scholar in Digital Humanities, in the Department of Computer Science, at Maynooth University. Her research focuses on the future(s) of web archive research in Ireland. Sharon holds a BA (Hons) in Cultural Studies, an MA in Digital Humanities, and a PG Diploma in Historical Archives. Sharon worked on several DH projects including Letters 1916-23 (Maynooth University), the Air Corps Aerial Photographs collection (Military Archives), and as a digital curator with TechArchives, Ireland. This study forms part of the doctoral research of Sharon Healy, which is kindly funded by the Irish Research Council under grant number [GOIPG/2018/1543].

**Helena Byrne** is the Curator of Web Archives at the British Library. She was the Lead Curator on the IIPC Content Development Group 2022, 2018 and 2016 Olympic and Paralympic collections. Helena completed a Master's in Library and Information Studies at University College Dublin, Ireland in 2015. Previously she worked as an English language teacher in Turkey, South Korea, and Ireland. Helena is also an independent researcher that focuses on the history of women's football in Ireland. Her previous publications cover both web archives and sports history.

**Katharina Schmid** is an IT developer at the Bavarian State Library. She holds an MA in European Literatures and Cultures and an MSc in Computer Science for Graduates in the Humanities or Social Sciences. Previously she contributed to a research project on applying methods from the digital humanities to web archives.

**Nicola Bingham** is Lead Curator of Web Archiving at the British Library and Co-Chair of the International Internet Preservation Consortium Content Development Group. She holds B.A and M.A degrees in History from the University of Newcastle upon Tyne and began her archival career at Tyne and Wear Archives Service before joining the British Library in 2002. She is responsible for web archiving strategy, ensuring that stakeholders across the Legal Deposit Libraries, and other partners, have the necessary tools for curating websites according to their own collection development policies.

**Olga Holownia** is the IIPC Senior Program Officer, based at Council on Library and Information Resources. Olga manages the communications and programmes of the International Internet Preservation Consortium (IIPC). Her key projects include the organisation of the annual IIPC General Assembly, Web Archiving Conference, training events and webinars. She has been responsible for research engagement and outreach activities and co-chairs the IIPC Research

Working Group. Olga has been involved in Digital Humanities projects since 2005 and she has served on the Board of the Digital Humanities in the Nordic and Baltic Countries since 2017.

**Michael Kurzmeier** is a recent PhD graduate in Digital Humanities and Media Studies at Maynooth University. His work revolves around the intersections of technology and society. His IRC-funded PhD thesis 'Political Expression in Web defacements' investigates political expression through hacking and introduces novel methods for retrieval and analysis of this special kind of archived web material.

**Robert Jansma** is a research associate and doctoral candidate at the University of Siegen. He is from a computer science background with a specialisation in software engineering and has put those skills to good use preserving and studying digital heritage. His research interests include commenting systems, digital methods, web histories, software histories and archiving methods for born digital material. Robert's research is kindly supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG): SFB-1472 - B03.

## Acknowledgements

We have many individuals and organisations to thank for their assistance, support, and encouragement throughout the duration of this project. First, we would like to thank the WARCnet Steering Group for organising the network meetings and activities, which enables members to meet like-minded people, and provides the capacity to develop projects such as this one. Next, we would like to thank Maynooth University, the British Library, the Bavarian State Library, the International Internet Preservation Consortium, and the University of Siegen for providing back bone support to the research team. We would like to acknowledge that some members of the research team are supported by the German Research Foundation (DFG), and the Irish Research Council (IRC). We are also very thankful to Dr Joseph Timoney (Maynooth University) and Professor Jane Winters (School of Advanced Study, University of London), who offered supervisorial assistance and support throughout. To add, we would like to express our gratitude and appreciation to the external reviewers who gave up their time to read the final drafts of this report and provide us with valuable feedback. Finally, we sincerely thank the respondents of this study, in giving us their time and sharing their experiences, for without them this study would not have been possible.

# Accessibility

As part of our commitment to accessibility, we have tried to ensure that the URLs provided in this document, in footnotes and the bibliography, are (i) captured in a web archive close to the time of access on the live web or (ii) saved by a member of the research team in a web archive close to the time of access on the live web. In case of future link rot, we have documented which archive the URL may be found in, both in the report bibliography and the project web library in Zotero. We use the [Extra] field in Zotero for this purpose, e.g. [URL Memento: Wayback Machine]. The resource is available as: Zotero | Groups > Skills, Tools, and Knowledge Ecologies in Web Archive Research.[1]

While it was not possible to capture the dynamics of the project web library in a web archive, we produced a full bibliography, with an accompanying dataset of bibliographic export files (e.g., BibTex, CSL JSON, CSV, etc.) which is available to download through the WARST Project files, available in Open Science Framework (https://osf.io/vf7gt/). Regarding paywall journal articles, we have attempted to provide a DOI, when available, and in the case of open access journal articles, we have further attempted to capture the source URL in a web archive, which we document in the [Extra] field in Zotero also.

The report is written in English, as it is a common language of the research team, however, we recognise that this may not be accessible for non-English speakers. To assist with this in some small way, we provide a translation of the Executive Summary, in Danish, French, Spanish and Catalan at the end of this document. We thank the WARCnet Steering Group for providing the financial assistance for translation services. We further thank Dr Juan-José Boté-Vericad of the University of Barcelona for providing the Catalan translation, and Jakob Moesgaard of the Royal Danish Library, and Ben Els from the National Library of Luxembourg for some last-minute proofreading.

To further assist with accessibility, we utilise the Arial font for headings, and the Calibri font in the body of the report. Arial and Calibri are part of the sans serif font family, which is a recommended family of fonts for web accessibility (Recite Me, n.d.). We also apply [alt text] for all images contained in this document. Should a reader need to access this document in some other form which would provide better accessibility, please contact the principal investigator, Sharon Healy.

---

1 https://www.zotero.org/groups/4669886/skills_tools_and_knowledge_ecologies_in_web_archive_research

# Glossary of Terms

In compiling this report, some members of the research team developed a glossary of terms and concepts which were used in the writing of the report, but also of terms which might be useful for novices starting out in web archive research. The glossary was further developed through a collaborative workshop organised by WARCnet working groups, WG3 and WG5, at the WARCnet London Meeting in June 2022. The glossary is published as a WARCnet Paper, titled 'Towards a Glossary for Web Archive Research: Version 1.0'.[2] The paper presents a glossary of more than 400 terms which have relevance for web archive research, and thus, it serves as a glossary companion for the current report. An interactive glossary resource is also available.[3]

# Abbreviations

| | |
|---|---|
| AADDA | Analytical Access to the Domain Dark Archive |
| AOIR | Association of Internet Researchers |
| AUT | Archives Unleashed Toolkit |
| BnF | Bibliothèque nationale de France |
| BUDDAH | Big UK Domain Data for the Arts and Humanities project |
| ccTLD | country code Top Level Domain |
| DH | Digital Humanities |
| DIGLIB | Digital Libraries Research Mailing List |
| DFG | German Research Foundation (Deutsche Forschungsgemeinschaft) |
| HAW | Croatian Web Archive |
| IIPC | International Internet Preservation Consortium |
| IFLA | International Federation of Library Associations |

[2] Healy, S., Byrne, H., Schmid, K., Floody, L., Boté-Vericad, J.-J. (2022) Towards a Glossary for Web Archive Research: Version 1.0. *WARCnet Papers*, September 2022. Aarhus, Denmark: WARCnet (ISSN 2597-0615)

[3] Zotero | Groups > Towards a Glossary for Web Archive Research. Zotero Groups, https://www.zotero.org/groups/4380600/towards_a_glossary_for_web_archive_research.

| | |
|---|---|
| INA | Institut Nationale de l'Audiovisuel (France) |
| IRC | Irish Research Council |
| IT | Information Technology |
| JISC | Joint Information Systems Committee |
| NLI Web Archive | National Library of Ireland Web Archive |
| NDSA | National Digital Stewardship Alliance |
| ODU WS-DL | Old Dominion University, Web Science and Digital Libraries Research Group |
| PANDORA | Preserving and Accessing Networked Documentary Resources of Australia |
| PRONI Web Archive | Public Records Office of Northern Ireland Web Archive |
| PWID URI | Uniform Resource Identifier for Persistent Web IDentifiers |
| RESAW | Research Infrastructure for the Study of Archived Web Materials |
| SAA | Society of American Archivists |
| UK | United Kingdom |
| UKWA | UK Web Archive |
| URI | Uniform Resource Identifier |
| URL | Uniform Resource Locator |
| US | United States |
| WAIL | Web Archiving Integration Layer |
| WARCnet | Web ARChive studies network researching web domains and events |
| WARST | Web Archives - Researcher Skills & Tools Survey |

# List of Contents

## List of figures

# List of tables

# EXECUTIVE SUMMARY

This study is part of a collaborative project by researchers from Maynooth University, the British Library, the International Internet Preservation Consortium, the Bavarian State Library, and the University of Siegen. The research team are all members of Web ARChive studies network researching web domains and events (WARCnet, warcnet.eu). WARCnet is funded by the Independent Research Fund Denmark | Humanities (grant no 9055-00005B).

The study focuses on individuals around the globe who participate in web archive research, in the context of web archiving, curation, and the use of web archives and archived web content for research or other purposes. As such, it is targeted at both creators and users of web archives. We consider web archive research to be representative of the processes and activities described in Archive-It's web archiving lifecycle model from appraisal, acquisition, and preservation, to replay, access, use and reuse (Bragg & Hannah, 2013). The study sought to identify and document the skills, tools, and knowledge required to achieve a broad range of goals within the web archiving lifecycle and to explore the challenges for participation in web archive research as well as the interludes of such challenges across communities of practice. We suggest that there is a perpetual need to examine the roles of skills, tools, and methods associated with the web archiving lifecycle as long as internet, web and software technologies keep advancing, upgrading, and changing.

The methodology for the study entailed desk research, participation in WARCnet meeting discussions, and an online questionnaire. The questionnaire was circulated via social media and email from 23 July to 21 September 2021. The recruitment strategy was to target archivists, librarians, curators, information managers, scholars, researchers, students, historians etc., and consisted of social media posts, and recruitment emails to network lists for archivists, librarians, curators, digital humanities, internet studies, and web archive studies. The results are based on a final number of 44 participants.

## Demographics

In this study, the participants (N=44) are aged between 18-64 years, and identify with residing in North America, Europe, and Asia. Participants identify with being at novice, intermediate and experienced levels for working with, or using web archives, and there is an equal representation of participants who identify with being male and female. This may provide some indication that gender does not present itself as an obvious barrier in web archive research, in this study at least.

Regarding the positional background of the participants, we offer two thematic representations being (i) participants who identified with working in a library, archive, or web archive environment (n=30), and (ii) participants who identified as being a scholar, academic, lecturer, post-grad/PhD student, or working in an IT/web design environment (n=14). Initially, we thought it would be possible to align participants' positions with whether they were creators of web archives, or users of web archives, but this was not the case. In fact, the boundaries were blurred as some respondents in the web archiving community also indicate that they are users of web archives as part of their work. While some respondents from the scholarly community indicate that they are creators/curators of web archives for research purposes. Thus, the categorisation of participants' positions was not as clear-cut as originally imagined, and we acknowledge that there is some overlap.

Broadly based on the participants' interests, backgrounds, experiences, and their relations to web archive research, we suggest that the participants in this study identify with one or more of the following subject areas, in alphabetical order:

- Arts, Humanities, Digital Humanities, Social Sciences, Media Studies
- Business and/or Law
- Data science/analysis, Statistics
- Information sciences (other than web archiving/curation)
- Internet/web applications, systems
- IT/Computer applications, systems, environments
- Use of web archives and archived web content
- Web archives, web archiving, curation

## Main Findings and Insights

In this summary we offer an overview of the findings and discussion, and organise it broadly into four main sections as follows:

- Skills, knowledge, tools, and methods in web archive research
- Challenges with web archive research
- Challenges with legal deposit, copyright, and GDPR
- Collaborations are key

## Skills, knowledge, tools, and methods in web archive research

From the findings, we presented a large array of skills, tools, methods, and knowledge which are required, desirable or useful for the domain of web archive research, across communities of practice. Some of the main representations include:

- Software and tools
- Web archives, web archiving, curation
- Programming, scripting languages
- Digital curation processes/workflows
- Data analysis skills
- Research methods/approaches
- Web design/internet related skills
- Information sciences (other than web archiving/curation)

The study shows several commonalities between participants who identified with working in a library, archive, or web archive environment, and participants who identified as being a scholar, academic, lecturer, student, or working in an IT/web design environment. For example, respondents from both communities indicate the use of web archives to find information, literature, and old websites, and show similar concerns about the losses and changes in web content. Dealing with exceptionally large volumes of data is further mentioned as a challenge for respondents from both communities. And respondents from both communities indicate the importance of acquiring knowledge and technical and critical skills through training, courses, and workshops, as well as through collaborations and mentorship. What also appears evident from various sections of the results, are the number of respondents from both communities who offer indications of the need for collaborations and pathways to develop further connections between creators/curators and users/researchers.

In terms of tools and methods, both communities would benefit from training in various capture methods including crawling software, screenshot, screen capture, and screencasting tools, and tools for downloading data from APIs. There are also indications that the development of training materials in the use of spreadsheet software, and the management and preservation of spreadsheets as data outputs would be useful for novice, intermediate and more advanced levels across the web archive research community as a whole. Furthermore, the study offers indications that users of web archives would benefit from introductory web archiving training, while staff in a web archiving environment would benefit from gaining some understanding and training in the tools and methods being utilised by user/researchers to analyse archived web data. Although, we should point out that the study

shows that participants from a scholarly or academic environment engage with a diversity of tools and methods. Moreover, the research question or methodology often influences which tools and methods are chosen, e.g., in cases when data is collected manually for close reading or when only specific parts of a website are scraped. This group of participants also face challenges due to a lack of research methods, theory, and approaches for combining traditional methods with web archive research. Thus, both communities would benefit from collaborative communal training in terms of current research approaches and methods for using the archived web, inclusive of demonstrations of tools and software. In this way, the field would be enriched through the input of dialogue by both communities in developing a better understanding of research methods and approaches for using web archives, as well as for "Gaining a proper understanding of archived web as a specific type of source and the consequences of these characteristics" for research using the archived web, as pointed out by one respondent.

## Challenges with web archive research

The study identifies multiple challenges which impact across the communities of practice. For example, challenges in capturing dynamic web content often result in archival deficiencies, which may further manifest as inconsistent and incomplete archival copies for the end user. Issues of incompleteness due to missing assets or broken links on live websites are problematic for both web archivists and end users, particularly when the gaps are hard to document and explain to users. The production of comprehensive metadata and documentation for web archive collections is an enormous challenge for archiving institutions as it is a time-consuming and labour-intensive process, exacerbated by the huge scale of the data. Less than complete metadata and documentation is then problematic for the end user seeking to engage with the collections. In addition, a lack of resources, and specialised skill sets may also affect the development of comprehensive documentation, which would facilitate the diversity of users, who further have different levels of skills and experience. There is also a need to consider that academic researchers and other end users such as journalists or lawyers may not have the time or energy to invest in acquiring a good comprehension of these issues, and thus, this may be perceived as a barrier to entry or challenge to engagement with web archives. Thus, there would be some benefit in providing users and potential users with introductory web archiving training, in a localised context relative to the web archive being used, in a bid to offer more awareness, and thus, more understanding of the scope of the collections vis-à-vis the limitations of archival strategies due to technical challenges, legal constraints, and lack of resources. It also presents an opportunity for collaboration between web archives and their users to develop

documentation in unison, which could eventually be tailored across disciplines and professions. This would be a significant gain for both communities creating a virtuous circle of creation and end use.

Challenges in learning new skills are experienced by respondents from both communities. We highlight how both communities would benefit from the provision of collaborative communal training across the full range of activities in the web archiving lifecycle. The study offers an overview of the types of skills and knowledge web archive practitioners and web archive users had prior to working with web archives, the skills they developed while working with web archives and the challenges they faced working with this type of resource. We propose that this might be used as a starting point to foster discussions in developing effective training materials for the necessary skills and tools for working with web archives across the spectrum of creator, curator, technician, or user/researcher. We further suggest that such training will also need to be benchmarked in a skills matrix, as it is very hard to develop and provide adequate training without a benchmark to measure against. We also find that the challenges experienced by the participants in the study do not become less with increasing experience and highlight the need for training across all levels of experience. We suggest that, in order to develop targeted resources for both introductory and more advanced training, further research would be required to see how challenges shift with increasing experience across communities.

## Challenges with legal deposit, copyright, and GDPR

Challenges with legalities, such as legal deposit, copyright, and GDPR present barriers for both the web archiving and researcher/user communities. Respondents from both groups discuss challenges for citing archived web content from legal deposit archives, or archives with restrictive access. Participants who identified with the web archiving community mention challenges in providing access to archived web collections due to legislation, copyright, GDPR, and embargoes. Challenges due to low response rates in acquiring permissions from website owners, are also mentioned, for both the capture of sites, as well as in providing access to the archived sites outside of a physical building. Further highlighted is the fact that while legal deposit may allow for the collection of websites by a legal deposit institution, it often does not effectively deal with the provision of access. For some institutions, access may only be provided onsite, which “makes them economically inaccessible” as noted by one respondent. This presents an area for more targeted research, as very little attention has been paid to the socio-economic factors which might influence barriers for entry and engagement with web archives.

Participants who identified with the academic community discuss challenges in using web archives due to legalities in terms of access to the data, use of the data, and storage of the data from web archives. Other challenges include handling copyright protected data from a web archive, as well as the inability to download data from some web archives. Challenges working on transnational collaborative projects are also found due to varying legal deposit laws across different countries which affect how the data is accessed, used, and by whom. Moreover, challenges in sharing data from web archives or making it reusable run counter to current trends by funders who increasingly stipulate open access and open science frameworks for research and data outputs. We suggest that further discussion and collaboration is required, to foster development in the application of research data management practises within legal deposit frameworks, open science frameworks, and web archive research environments. As a starting point there would be some benefit in providing introductory training and courses regarding (non-print) digital legal deposit for novices from both communities.

## Collaborations are key

Finally, the study finds positive acknowledgements which reinforces the need and the value of collaborations across communities of practice, and especially how such collaborations benefit both communities in addressing some of the challenges mentioned above. However, we must acknowledge that web archiving organisations and institutions may not have the resources to provide the necessary support for researchers. Reasons for this are varied and may be "due to a mix of curatorial, technical, legal, economic and organisational constraints" (Brügger, 2021, p. 217). Such factors may be further influenced by the political and economic climates in particular countries which may not be favourable to funding cultural heritage projects, or due a lack of capacity of web archiving to promote the value of web archives to stakeholders (i.e., through user case studies). Indeed, this presents a paradox, whereby web archiving organisations need resources to assist researchers to develop user case studies to demonstrate the value of web archives to attain funding to provide support to researchers. Thus, for organisations who wish to seek funding to develop web archiving initiatives it is imperative to make a business case (from the outset) for activities in the full web archiving life cycle, inclusive of providing access and support mechanisms for academic researchers, or other end users such as journalists or lawyers.

# 1. INTRODUCTION

Web Archives – Researcher Skills & Tools Survey (WARST) is a collaborative project by researchers from Maynooth University, the British Library, the International Internet Preservation Consortium, the Bavarian State Library, and the University of Siegen. Sharon Healy (Maynooth University) acted as the principal investigator for the project, and it received ethics approval (SRESC-2021-2436150). The research team are all members of WARCnet, and between them, have backgrounds in humanities, digital humanities, cultural studies, media studies, cultural heritage, library and information science, archival science, computer science, and IT development.

Several talks and activities at the WARCnet networking meetings (2020-2021) highlighted the need to examine the roles of skills, tools, and knowledge for conducting web archive research. Web ARChive studies network researching web domains and events (WARCnet) is a transnational interdisciplinary network, primarily based in Europe. It provides network meetings and activities for web archivists, IT developers and researchers who study the archived web, with the involvement of some leading European web archives, and the International Internet Preservation Consortium (IIPC) (Brügger, 2020; WARCnet, About WARCnet, n.d.). WARCnet is funded by the Independent Research Fund Denmark | Humanities (grant no 9055-00005B). From the meetings, it soon became clear that web archiving and curation, as well as the use of the archived web for research or other purposes, comes with its own set of social, cultural, geographical, legal, ethical, institutional, and technical challenges. Moreover, the creation and use of web archives continually evolves due to the rapid advancements in internet, web, and software technologies. Hence, this prompted further interest to investigate some of the effects of these challenges, in line with skills, tools and knowledge. Consequently, the concept for the WARST project was realised.

This study explores the skills, tools, and knowledge ecologies in web archive research, and focuses on individuals around the globe who participate in web archiving, curation, and the use of web archives and archived web content for research or other purposes. We further consider web archive research to be representative of the processes and activities described in the Archive-It's web archiving lifecycle model from appraisal, acquisition, and preservation, to replay, access, use and reuse (Bragg & Hannah, 2013).

## 1.1 Background

While the web was founded in the early 1990s, on the principles of sharing information between scientists, it rapidly became a space for more diversified forms of information as internet technology advanced and became more affordable (Masanès, 2006, p. 3). By 1997, an article in Time Magazine hailed that the "World Wide Web could prove as important as the printing press" (Wright, 1997). Indeed, by the early 2000s, the web was claimed to be "the information source of first resort for millions of readers" (Lyman, 2002, p. 38). Meanwhile, concerns about the transient nature of content on the web also emerged due to invalid and broken links, also known as link rot, link decay, or reference rot. Alexa Internet statistics from 1998 estimated that the web was "growing at the rate of 1.5 million pages daily", but that 1% of these web pages also disappeared after a week (Library of Congress, Public Affairs Office, 1998; Quint, 1998). Indeed, several studies have been conducted across numerous disciplines which examine the transience of the web through studies on link rot and web content drift, website evolution, or deletion (Harter & Kim, 1996; Lawrence & Giles, 1999; Koehler, 1999; Kitchens & Mosley, 2000; Cho & Garcia-Molina, 2000; Germain, 2002, Fetterly et al., 2003). For example, a study by Spinellis (2003) used two computer science journals to source a sampling of publications from 1995-1999 which cited URL (Uniform Resource Locator) references. They extracted 4,375 URL references for verification and suggest that 20% of URLs were inaccessible after one year of publication, and that this increased from 40% to 50% four years after publication. Spinellis (2003) argues: "Citations in scholarly work are used to build upon existing work [therefore] references that cannot be located seriously undermine the foundations of modern scientific discourse" (p. 71).

From at least 1994, libraries, archives and cultural heritage organisations have also had concerns about the ephemerality of web content.[4] At the same time, the development of web crawler programmes gave rise to the technology for web archiving (Schneider et al., 2009, p. 206). Some of the pioneering efforts in web archiving may be attributable to the National Library of Canada from 1995, the Internet Archive from 1996, the National Library of Australia (PANDORA) from 1996, and the Royal Library of Sweden conducted their first Swedish domain crawl in the summer of 1997 (Webster, 2017, pp. 176–178; Koerbin, 2021, p. 24; Arvidson et al., 2000). A study by Gomes et al. (2011) observes a significant growth in web archiving

[4] The National Library of Canada (now part of Library and Archives Canada) initiated discussions in 1994 around the collection of electronic materials, inclusive of websites; and initiated a pilot project in 1995 (Webster, 2017, p. 177). The National Library of Australia organised a working group to address collection and archiving techniques for the Web in 1995 and initiated a web archiving programme in 1996 (Schneider et al., p. 206).

initiatives from 2003, but mostly in developed countries. Their research also provided the base for a Wikipedia article which has seen a continual growth in web archiving activities by national libraries, heritage organisations and associations, academic institutions, and by non-profit and commercial organisations (Wikipedia, 2011+, List of Web archiving initiatives). Immediately evident, is a strong representation of web archiving initiatives in Europe and North America.

It is widely agreed that web archiving involves selection and collection of web content, preserving it for the future, and making it available for access and use (Niu, 2012; IIPC, Web Archiving, n.d.). According to Niu (2012), the library/archive communities tend to refer to appraisal, as "the process of evaluating the value of records and deciding whether and how long records should be preserved. It is essentially a process of selection." The process of selecting web content for archival purposes involves many variables, but in general it tends to be organised around a domain type or name, a topic or event, a media type or genre (Niu, 2012; Hockx-Yu, 2011). Masanès (2005, p. 75) describes this as "site-, topic-, or domain-centric" selection. Archiving based on media type such as online newspapers, or genre such as video games, already has some primary boundaries for selection criteria. However, archiving based on a topic or event tends to depend on human assessment for identification within the selection process (Niu, 2012). Selections based on a domain type or name (e.g., .com, .org, .net) or by a country code Top Level Domain (ccTLD) (e.g., .fr, .ie, .de) might be easily automated, and may be necessitated by national laws such as legal deposit legislation (Masanès, 2005, pp. 75–76; Hockx-Yu, 2011, pp. 1–2). Social media archiving also comes under the umbrella of web archiving and may involve a different set of workflows and archival tools in comparison to archiving a static or semi-static web page, as well as different legal, ethical, and curatorial considerations (Breed, 2019; Michel et al., 2021; Vlassenroot et al., 2021).

Web curation tends to set the guidelines, rules, and procedures for selecting and collecting web content and ensuring that the web content matches the "curatorial objectives" (Schneider et al., 2009, pp. 210–11). For example, this may involve the development of collection policies, determining the scope for a legal deposit crawl, or making decisions on whether to pursue permissions for external web pages for a selective thematic collection. In some cases, permissions may also need to be sought by an institution which chooses to archive content outside of a national domain for instance, and/or to provide access to content outside of a reading room, as is the case for national libraries in Estonia, New Zealand, and the United Kingdom (UK) (IIPC, Legal Deposit, n.d.; Byrne, 2020).

Historically many countries have enacted legal deposit regulations, which in general terms, mandate that publishers who operate within the national domain are legally required to deposit at least one copy of each publication in a nominated institution, often designated as the national library. Legal deposit serves as a system to compile, maintain, and provide access to a comprehensive collection and bibliographic record of a country's published output and, in doing so, creates a significant manifestation of national cultural heritage. In embracing and addressing the digital age, several countries have amended their legal deposit legislation to incorporate the deposit of non-print materials such as electronic publications stored on devices like CD-ROMs or published online, as well as the archiving of national web domains at scale. In Denmark, for instance, legal deposit legislation for print publications has existed since 1697, and legislation for the legal deposit of non-print materials was introduced in 1997 (Dupont, 1999, pp. 244-245). The legislation was further revised in 2004 to broaden the scope for the inclusion of archiving the Danish national web domain (Webster, 2017, pp. 179–180). In the UK, legal deposit has been a part of English law since 1662 for printed publications (The Bodleian Libraries, Legal Deposit, n.d.) but it was not until April 2013 when Non-Print Legal Deposit Regulations were enacted, which enabled the archiving of the UK national domain. Web archiving is therefore a legal obligation on the part of some, but not all legal deposit institutions. For example, in Ireland digital legal deposit was enacted through the Copyright and Other Intellectual Property Law Provisions Act 2019, which allows for the collection of e-books and online journals, however, the legislation does not allow for the archiving of the Irish national web domain (Ryan et al., 2022).

While legislation in some countries has allowed the collection and preservation of web content, access to this content varies widely. For instance, legal deposit legislation often mandates that materials collected under the auspices of the regulations may only be accessed on the premises of the legal deposit library(ies). This makes sense considering the library is charged with conserving this material for the benefit of future generations. However, this clause has often been extended to archived web materials, creating a paradox whereby archived websites, which were originally published, and publicly available on the web, may only be accessed on terminals in library reading rooms. Moreover, access to web archives varies from country to country. For example, national web archives such as the Croatian web archive and Icelandic web archive are completely open access, the UK web archive and New Zealand web archive are a mix of open access and onsite access, the French web archive is onsite only, the Danish web archive can be accessed offsite by legitimate researchers on a project permissions basis, while access to the Swedish web archive is prohibited by the Swedish Legal Deposit (IIPC, Legal Deposit, n.d.; Winters, 2020a, pp. 160–163).

In terms of preservation, Day (2006) describes web content preservation as a subset of digital preservation which is concerned with the processes of maintaining captured web content in a usable and accessible condition for the long-term. Web preservation may also be concerned with web archaeology. In explaining web archaeology Tjarda de Haan notes the following:

> Data is the new clay, scripts are the new spades and the World Wide Web is the youngest layer that we are digging up. Web archaeology is a new area in e-culture where we excavate and reconstruct relatively new (born-digital) objects, which were lost not so long ago, using new digital tools. Both the archaeological finds and the methods of unearthing and reconstructing our digital past are very recent and still in development (de Haan, 2018).

Other commentators suggest that where possible websites should not only be preserved as web archives but also as the software itself, preserving the dynamic nature of the website (Alberts et al., 2017; De Haan et al., 2017). Dynamic preservation of the software opens new research opportunities while raising new challenges for preservation, such as keeping the software functional and accessible across time and systems.

The value of web archives as resources has also received some attention. Gomez and Costa (2014) offer an overview on the importance of web archives in the humanities for current and future historical research. Milligan (2019) exhibits the value of web archives for historians, using computational tools, to analyse websites from GeoCities. Developed from the mid-1990s, GeoCities was a free web hosting platform which had more than 2 million members by the time it was bought by Yahoo in 1999 (Mackinnon, 2022). Such websites are often only examinable through a web archive, as for the most part, GeoCities was taken offline when Yahoo discontinued the service in 2009 (McKinnon, 2022; Shankland, 2009). For some reason, GeoCities Japan (GeoCities.co.jp) escaped the 2009 closure, until Yahoo finally announced its closure for the end of March 2019 (Archiveteam, 2018+; Gottsegen, 2018). Winters (2017) draws attention to the use of web archives by news and media outlets, to highlight the disappearance of web content such as political party documents, and political campaign websites, while Healy (2019) notes the benefits of web archives for studying Irish LGBT history. Gorsky (2015) discusses the value of web archives for examining contemporary public health, while Adelmann and Franken (2020) discuss the value of archiving the web for studying telemedicine within digital health systems. Kurzmeier (2020) demonstrates the value of web archives for the study of political communication through hacked websites, while Huc-Hepher and Wells (2021) offer a discussion on the use of diasporic web collections in a web archive for studying histories of migrant communities in London. The value of web archiving has also rippled into business and law. For example, Costa and Silva (2010) suggest that web archives provide a resource for use cases to develop company trustability profiles.

Denev et al. (2010) discuss how web archiving is of benefit for business and market analysts, for legal experts on intellectual property and internet compliance, and for investigating internet fraud and consumer rights violations. And Eltgroth (2009) and Taylor (2017) examine the use of archived web content as evidence in a court of law.

Unlike other traditional forms of information that humans interact with, the web is an ever-changing space for new, old, updated, and deleted content. Thus, the rationale for archiving the web entails that it is necessary to record and preserve a fleeting cultural, historical, evidential, informational, and social record, as well as to provide a means for access, research, and analysis. While this may seem straightforward, web archiving and curation is a complicated process requiring constant decision making (Lyman, 2002; Dougherty, 2007, p. 19). It requires decisions on the appraisal and selection of content to be captured (Summers, 2020; Post, 2017; Summers & Punzalan, 2017). Lyman (2002) suggests that decisions need to be made about authenticity and provenance in order to define "the boundaries of the object to be collected" (p. 42). The Society of American Archivists (SAA) Dictionary of Archives Terminology (2005+), offers a definition of provenance as: (i) "the origin or source of something", and (ii) "information regarding the origins, custody, and ownership of an item or collection." Further decisions need to be made on the technology to be used for permission management, as well as for capture and replay (Grotke & Jones, 2010; Xie et al., 2013; Bingham & Byrne, 2021; Jackson, 2022). With more decisions to be made on how to make the data accessible for use—and to whom?—which may also coincide with a set of legal requirements (Jacobsen, 2008; Hockx-Yu, 2014; Winters, 2020a). Decisions regarding "the ethics of archiving the web" are also highlighted by Graham (2019), and raises the question of "How does this type of collecting fit into existing ethics of collecting and where does it demand that we develop new practices and principles?" (p. 103). Moreover, such decisions will also depend on the availability of resources, as well as organisational IT infrastructures (Anthony, 2013; Post, 2017; Brügger, 2021). Summing this up succinctly, Vlassenroot et. al (2019) suggests that "web archiving requires a strategic approach as much is required in terms of technologies, systems, policies, procedures and resources to make web archiving more than merely harvesting and storing online content" (p. 86).

Web archiving is further complicated by "ever-evolving" web and internet technologies. As Truman (2016) points out: "the ever-evolving nature of the web means that the live Web and Internet technology will always be ahead of the capture tools" (p. 20). So, as a process, web archiving also relates to research on crawler-based archiving, techniques for improving crawler efficiency to enable better data quality assurances, techniques for examining data quality of a web archive, or examining quality metrics (Denev et al., 2009; Spaniol et al., 2009;

Denev et al., 2011; Xie et al., 2013; Bingham, 2014). And all this will be accompanied by continual research on techniques and software developments for search and retrieval, replay/playback, digital preservation, software archaeology, IT integrations, and more (Mourão & Gomes, 2021; Newing & Clegg, 2021; Samar et al. 2017; Jackson, 2022; UK Web Archive, 2018; CCSDS-DAI, 2021; Alberts et al., 2017; Jansma, 2020; Beis et al., 2019). Therefore, the web archiving life cycle of tools will keep changing too. By this, we refer to Truman's (2016) description as "tools that have been developed to address various functional needs across the lifecycle of web archiving (from capture to access and analysis by researchers)" (p. 7). Thus, we also consider the development and implementation of software and tools as a necessary part of the activities and processes undertaken in web archive research.

Engagement with web archives for scholarly research purposes has also developed in the past decade or so (Maemura, 2022), and is evident in the accumulation of literature published in edited collections in recent years (Gomes et al., 2021; Brügger & Laursen, 2019; Brügger & Milligan, 2019; Brügger, 2017; Brügger & Schroeder, 2017). However, several commentators observe how scholars were slow to engage with web archives as a research resource (Webster, 2020; Webster, 2017; Winters, 2017; Leetaru, 2017; Meyer et al., 2011; Dougherty et al., 2010). Indeed, Meyer et al. (2011) were of the opinion that "the use cases for web archives are not well articulated and have not engaged the research community in any significant way" (p. 4). There are several reasons put forward for the slow development of researcher engagement with web archives. Obvious reasons include a lack of awareness, or simply because some academic disciplines have no need to rely on such sources (Jatowt, 2008; Riley & Crookston, 2015; Costea, 2018; Healy, 2021).

Other reasons relate to the challenges in understanding the characteristics of the archived web, as an archived website or web page is almost never a complete copy or surrogate of what was once on the live web (Brügger, 2010, p. 6; Brügger & Finnemann, 2013, p. 74). Indeed, Brügger and Finnemann (2013) propose: "The Archived Web is a Reborn, Unique and Deficient Version and Not Simply a Copy of What was Once Online" (p. 74). For example, there are limitations with the software/hardware to capture some types of dynamic content which may result in deficiencies (Brügger, 2010, p. 6; Pennock, 2013, p. 13; Bingham and Byrne, 2021, p. 2; Jackson et al., 2016, p. 103). Other deficiencies may occur due to the time it takes to capture, and the fact that some content may be updated during capture (Brügger, 2010, p. 7). Other commentators note challenges due to the variances between searching on the live web, and searching in a web archive (Costa, 2021; Helzmann & Nejdl, 2021; Winters & Prescott, 2019; Jackson et al., 2016; Nielsen, 2016). Most web archives offer URL search as an

entry point to find archived web materials, such as the Internet Archive's Wayback Machine, the UK Web Archive, and Arquivo.pt (the Portuguese web archive), however the user would need to know the URL in the first place. A few web archives, allow for alphabetical browsing such as the UK Government Web Archive, or browsing through topical collections, such as the UK Web Archive and the BnF Archives de l'internet (Vlassenroot et al., 2019, pp. 99-100). Several web archives allow for a full-text search. For some scholars, this is problematic for large web archives due to the enormous amount of query returns (Winters & Prescott, 2019, pp. 398–399; Jackson et al., 2016, p. 105; Nielsen, 2016, pp. 22–23). In addition, full-text searching within a web archive does not provide the same experience of search, or the behaviours of ranking we experience on the live web with search engines such as Google or Bing (Winters & Prescott, 2019, p. 398).

In fact, while the web archiving community has worked on improving its search capabilities by complementing traditional URL search with metadata and full-text search, they have encountered significant challenges along the way. As pointed out by Costa (2021) the "manual creation of metadata describing curated collections and their artefacts is a time-consuming and expensive process, which makes it a non-viable option for large-scale archives" (p. 72). Therefore, in the case of large web archives, Costa (2021) notes that metadata needs to be created automatically, which is a method used by up to 72% of the web archives around the world (p. 72). Setting up full-text search for text in a variety of different languages and file formats and building a search system that scales well across large collections is also a complex endeavour (Costa 2021, pp. 79–82). As users have pointed out, the potentially very large number of search results further requires an efficient ranking algorithm, for which there is no given solution. Algorithms that were developed for ranking search results from the live web will not provide satisfactory results, because web archive search also includes a temporal dimension. Collections typically include different versions of a given document and users are not necessarily interested in the latest one (Costa and Silva, 2010). Jackson et al. (2016) have also highlighted the difficulties of designing a ranking model that satisfies scholarly requirements, as "some scholars [...] questioned the very idea of relevance ranking" (p. 105). If a ranking model is used, Jackson et al. (2016) argue, it must be made completely transparent so scholars can interpret the results accordingly.

For some commentators a lack of research engagement is due to a lack of collaboration and communications between web archiving initiatives, and users/researchers. For example, in a Harvard Library report, Truman (2016, p. 3) identifies the need for more communication and collaboration between the creators of web archives and those who use web archives for research, as well as potential users. One also needs to consider here that, initially, web

archiving organisations did not have a priority on how their collections would be used, as part of their initial web archiving strategies. Rather, the main priority was to keep up with changing technologies to enable collection in the first instance (Dougherty et al., 2010, p. 10; Hockx-Yu, 2014, p. 113; Webster, 2017, p. 187; Huurdeman & Kamps, 2017). It is also worth noting that not all web archives are able to support researcher engagement "due to a mix of curatorial, technical, legal, economic and organisational constraints" (Brügger, 2021, p. 217). However, in recent years collaborations between web archives and researchers have greatly improved (Schroeder & Brügger, 2017, pp. 12–13; Winters, 2020a, p. 169). In part, this can be attributed to growing efforts by consortiums, networks, and research projects to develop collaborations to increase research engagement. Some of which include the International Internet Preservation Consortium (IIPC), the Research Infrastructure for the Study of Archived Web Materials (RESAW), the Big UK Domain Data for the Arts and Humanities project (BUDDAH), the Web90 project (Web90: Heritage, Memory and History of the Web of the 1990s); the Web Science and Digital Libraries Research Group at Old Dominion University (ODU WS-DL), the Web ARChive studies network researching web domains and events (WARCnet), ResPaDon (network to develop and diversify the uses of web archives) and the Archives Unleashed project. The Archives Unleashed project have also launched a program for cohorts, to further foster research engagement.

Other studies have also looked at challenges for researcher engagement. For Truman (2016), challenges arise for researchers due to a lack of technical knowledge in the application of data mining techniques to vast volumes of data, as well as a lack of training and experience in using web archives, from discovery processes to integrating the use of archived web content with traditional research approaches. Costea (2018) identifies a need for improvements to web archives to satisfy researchers' needs in the areas of discoverability options, data selection, data management, as well as access to more comprehensive documentation and metadata. Another requirement suggested by Costea (2018) is the ability for researchers to extract data from a web archive to create a dataset for their own research requirements. In another study on academic engagement. Healy (2021) also highlights challenges for researchers due to difficulties with search, navigation, and discovery functions of a web archive, as well as challenges in working with large volumes of data. Other challenges include a lack of understanding of copyright implications for using archived web content, and a lack of understanding of what does or does not get captured in a web archive, and why (Healy, 2021).

Challenges for researchers also arise due to ethical issues. Graham (2019) argues that there has been little attention paid to "the ethics of experiencing and accessing the past web" (p. 103). For example, Graham (2019) highlights ethical challenges regarding biases, and reminds

us that “on the live web, biases are embedded into both the content and the discovery processes” of what is being collected by web archives. Therefore, Graham (2019) asks how web archivists are “replicating and/or intervening in how biases operate?” once web content is collected and moved “into the more fixed platform of the web archive” (p. 104). While noting the value of web archives as resources for researching online communities and bottom-up histories, Mackinnon (2021) also warns researchers of “significant ethical, methodological and epistemological issues” when it comes to the study of websites of “young people of the past” (pp. 442–443). Here, Mackinnon (2021) refers to the websites created by young people under the age of 18, which were once hosted on the free hosting GeoCities platform from the 1990s-2000s and ended up in a web archive due to the collection efforts of the Internet Archive when Yahoo announced the forthcoming closure of GeoCities in 2009. For Mackinnon (2021), this presents researchers with “opportunities for harmful data practices” while it also brings into the debate an “individuals’ ‘right to be forgotten’” (p. 442). Therefore, for Mackinnon (2021), researchers need “to consider whose stories are being told, who is equipped to tell them, and what kinds of vulnerability and harm one might encounter and create when doing so” (p. 443). Maemura (2018) also points to challenges due to “ethical implications of how materials are used”, as well as “questions of consent” and the responsibility of the researcher to the people represented in the data (p. 331).

Other challenges arise for researchers due to legalities, copyright and GDPR. For example, Winters (2020a) and Milligan (2015) discuss the challenges in using legal deposit collections which are only accessible on a library terminal in a designated reading room. Using the UK Web Archive legal deposit collections as an example, Winters (2020a) describes the locked down nature of the library terminal for accessing/viewing the captured websites, and how “no two people in the same legal deposit library can simultaneously view the same instance of a captured web page” (p. 164). Moreover, due to legal deposit restrictions users cannot take a screenshot of a captured web page, a photograph of the screen, which would otherwise be allowed for historians viewing print documents in a reading room (Milligan, 2015). Nor can they view the source code, which is an object of study by itself (Milligan, 2015). Truter (2021) further highlights how researchers using web archives encounter challenges in the access and use of archived web data/materials due to legal restrictions, inclusive of copyright and third-party ownership, privacy policies, and the General Data Protection Regulation (GDPR) in the European Union (EU). Truter (2021) suggests that this creates challenges not only for the use of data from web archives but may also affect the ability to share the data or make it reusable.

Thus, as one can see, there are a multitude of challenges when it comes to archiving the web, as well as a multitude of challenges for those wishing to use the archived web for research or other purposes. This study seeks to explore such challenges, and their interludes across communities of practice.

## 1.2 Purpose of the Study

Web archiving has been around for a quarter of a century, and for some commentators, it may be seen as a field that is starting to mature beyond the establishment phase (Schafer & Winters, 2021, p. 130). In contrast the use of archived web materials for research or other purposes is much less established, with it only seeing progress in the past decade, or so (Maemura, 2022). As web archive research is still recognised as an emerging field of study, it is also difficult to define (Reyes Ayala, 2013; p. 1; Vlassenroot et al., 2019, p. 86). Thus, coming up with a universal definition for web archive research is not our goal. However, we do need some understanding of the extent and boundaries of web archive research.

Maemura (2018) offers a useful starting point in understanding the scope of web archive research and refers to web archive research as a broad term "to encompass the study of all activities involving web archives" (p. 327). Maemura (2018) offers several examples of such activities as follows:

- the creation of web archives
- the study of activities such as how collections are created with technical tools and systems like web crawlers,
- the organisational/curatorial aspects of collection development,
- the study of activities to support the use of web archives, through developing access interfaces, or specific research methods and techniques (p. 327).

Maemura (2018) also includes research which is related to:

- exploring, organising, and delimiting a corpus for study,
- critically examining collected materials,
- considerations for ethics, consent and responsibility of a researcher when using the archived web for scholarly purposes (p. 327).

Maemura's (2018) broad description of web archive research as "the study of all activities involving web archives", fits well for the purpose of this study. However, we also regard web archive research to be representative of the processes and activities described in the Archive-It's web archiving lifecycle model (Figure 1.1) from appraisal, acquisition, and preservation, to replay, access, use and reuse (Bragg & Hannah, 2013).

The Archives unleashed team note that scholarly work with web archives tends to follow the FEAV process model (Ruest et al., 2020). This descriptive model breaks down scholarly activity with web archives into four steps: Filter, Extract, Aggregate and Visualise. The filter step is to reduce the web archive to a more manageable subset. Filtering can be performed based on content, metadata, or extracted information. The extract step extracts specific information of interest to the researcher from the subset created in the filter step. Examples include the text from web pages, links, named entities and specific file types. The aggregate step aggregates or summarises the results of the previous two steps as a derivative of the original archive. Examples include counts, minimums, maximums, and averages. The visualise step visualises the aggregate for the scholars' consumption and interpretation. Examples would be tables, charts, and graphs. The derivatives from the aggregate step can be reused as input for a new iteration of the FEAV process model and can serve as a stepping stone to working with web archive data without needing to know how web archives themselves work (Ruest et al., 2020).

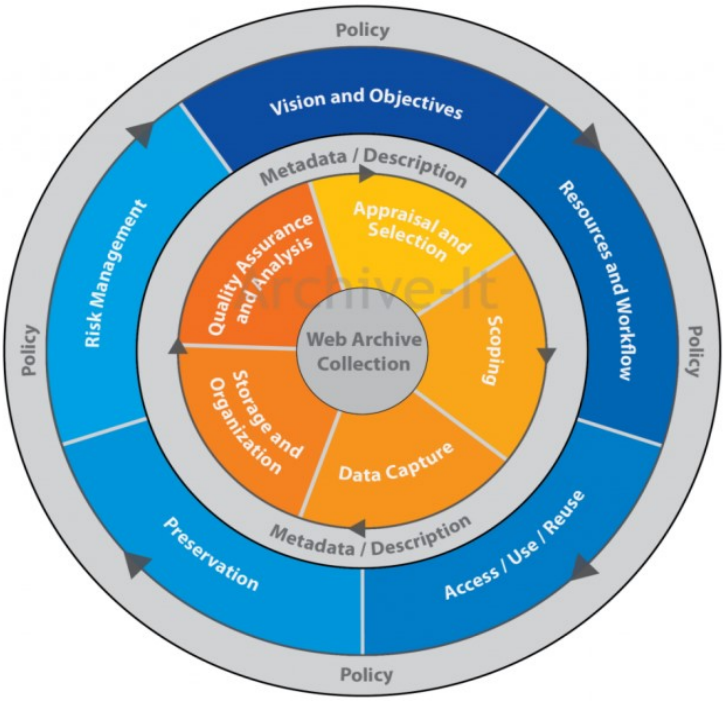


Figure 1.1: Web Archiving Life Cycle Model (Bragg & Hanna, 2013)

Thus, for the purpose of this study we consider web archive research to be inclusive of web archiving and curation, and the use of web archives, archived web content and their derivatives for research or other purposes. It is inclusive of the processes and activities throughout the web archiving lifecycle. We further maintain that as long as internet, web and

software technologies keep advancing, upgrading, and changing, there will always be a need to keep examining the roles of skills, tools, and methods associated with the web archiving lifecycle. Moreover, the circumstances (legal, ethical, curatorial, financial, technical, temporal, social, and political) under which an organisation (or individual) archives web collections, will also affect how such collections can be accessed, used, and interpreted by researchers and end users (Ben-David, 2021; Brügger, 2021; Ogden, Halford & Carr, 2017; Ogden, 2021; Vlassenroot et al, 2019). Therefore, we assert, there will always be a need to keep evaluating skills, tools, and knowledge ecologies for conducting web archive research across communities from creators to end users.

The focus of the study is individuals around the globe who participate in web archive research, in the context of web archiving, curation, and the use of web archives and archived web content for research or other purposes. The study seeks to identify, and document skills, tools and knowledge required to achieve a range of different research goals within the web archiving lifecycle and explores the challenges for participation in web archive research, and the interludes of such challenges across communities of practice.

Therefore, this report aims to:

- offer an overview of the skills, tools, and knowledge ecologies within web archive research,
- explore the challenges for the creation and use of web archives, and examine how these challenges interlude across communities of practice, and
- foster a discussion on the development of future training materials for the web archive research community.

In doing so, we are guided by the following research questions:

- What type of skills and knowledge are useful or important for conducting web archive research?
- What type of software, tools and methods are currently being used in web archive research?
- What are the main challenges for participation in web archive research?

## 1.3 Document Outline

In the next sections, we offer an overview of related literature and discuss our methodology. We then present the findings, which is organised into several sections as follows:

- Demographics
- Data, Tools & Methods

- Skills and Knowledge
- Citation Practises
- Resources and Data Sharing

Thereafter, we provide a discussion of seven main dimensions as follows:

- Participants - Positions, Backgrounds, and Interests
- Pathways to Web Archive Research
- Skills and Knowledge Ecologies in Web Archive Research
- Challenges with Web Archive Research
- Referencing the Archived Web and Data Sharing
- Software, Tools, and Methods used in Web Archive Research
- Challenges with Legal Deposit, Copyright, and GDPR

# 2. RELATED LITERATURE

This study has several overlaps with other web archive user and researcher engagement studies (Costa & Silva, 2010; Jatowt et al., 2008; Ras & van Bussel, 2008; Hockx-Yu, 2014; Riley & Crookston, 2015; Costea, 2018; Moiraghi, 2018; Healy, 2021). However, this study is not focused on one organisation, or indeed one country. Rather, it focuses on individuals around the globe, who have a relationship with web archiving and curation, and/or the use of the archived web for research, or other purposes. Thus, this research also has areas of overlap with studies focusing on web archiving practices and organisational structures (NDSA Content Working Group, 2012; Bailey et al. 2014; Bailey et al. 2017; Farrell et al. 2018). While we borrow from these studies, we also build on the work of Thomas et al. (2010), Dougherty et al. (2010), Truman (2016) and Vlassenroot et al. (2019). Such studies investigate the practises of international web archiving initiatives, as well as addressing the challenges for the use of web archives for research. Also, worth noting here is Truter's (2021) study which looks at research data management and sharing practices of researchers in web archive studies. In the next section we offer a review of a selection of these studies. To note here, we only examined literature in English, as it is the common language of the research team.

## 2.1 Web Archiving Tools & Services

The National Digital Stewardship Alliance (NDSA) conducted web archiving surveys in 2011, 2013, 2016, and 2017 which were, more or less, aimed to get a better understanding of the types of web archiving activities being conducted in the United States, the history and scope of such activities, the types of content being selected for preservation, the types of tools and services being used, the types of access and discovery options being provided, the types of permissions being sought for collection and access, and the types of policies in general operation across organisations (NDSA Content Working Group, 2012; Bailey et al., 2014; Bailey et al., 2017; Farrell et al., 2018). Founded in 2010, the NDSA is a voluntary organisation made up of a consortium of educational, governmental, non-profit, and commercial organisations committed to the long-term preservation of digital information (Farrell et al., 2018, p. 4). While we do not have the space here to explore each aim, we will focus on the findings in relation to the type of tools and services being used across web archiving organisations.

From the survey conducted in 2011 (N=72), 63 participants responded to the question on the use of tools/services for harvesting web content, of which 60% (=38) used an external service

for acquisition, 26% (=16) used an in-house method, and 14% (=9) used both an in-house method, and external services. A further 25 respondents provided details of their tools/software used for in-house crawling or in conjunction with an external service. Both Heritrix (24%, =6) and HTTrack (24%, =6) were most popular amongst the 25 respondents, followed by Wget (12%, =3), Teleport Pro (12%, =3) Adobe Web Capture (12%, =3), and Grab-a-Site (8%, =2) (NDSA Content Working Group, 2012). The 2013 survey (N=92) saw a slight increase in the number of organisations using external services, and a slight decrease in those using in-house crawling methods exclusively, with several organisations opting for both in-house methods, and external services. In terms of in-house harvesting methods, the study further indicates the use of Heritrix (29%) as the most popular crawler, followed by HTTrack (18%), Teleport Pro (9%), and Wget (7%). To note here, the study does not reveal the amount of participants who responded to this question, thus, it is difficult to get a feel for an amount in numbers, through percentages alone. Additionally, a high number of respondents (31%) provided other options regarding the use of in-house tools such as: modified versions of Heritrix, manual download of individual web files, screenshots, Social Feed Manager, tools for link extraction such as UXTR: Universal Links Extractor, and web archiving platforms such as KEN (Bailey et al, 2014, p. 18).

The 2016 survey (N=104), saw another increase in the use of external service providers, and an increase in the use of both external services, and in-house archiving methods, suggesting an increase in local experimentation with mixed approaches (Bailey et al. 2017, p. 23). Of 29 participants who answered the question on tools for in-house archiving, Heritrix (31%, =9) and HTTrack (28%, =8) were again the most popular tools, and the use of Webrecorder (21%, =6), surfaced as a new tool in 2016. Other tools mentioned include Adobe Web Capture, Brozzler, Grab-a-site, Teleport Pro, Wget, Umbra, WAIL, and the Web Curator Tool (Bailey et al. 2017, p. 23). The 2017 survey (N=119), saw a majority of institutions using external services for harvesting web materials, but also an increase in the number of institutions capturing web materials in-house, suggesting the dominance of external services as a method for institutions to conduct web archiving (Farrell et al. 2018). However, there was also a steady rate of increase in local capacities for in-house web archiving. Regarding the question on tools used for capturing web content, of 45 respondents who answered, Heritrix was shown as the most popular used tool, but also showed a decline in the use of HTTrack, which was popular in previous surveys, and a decline in tools such as Wget and Adobe Web Capture. While other tools mentioned in prior surveys, such as Grab-a-site, Teleporter Pro, and WAIL were not mentioned at all in this survey. On the other hand, this survey indicates “an explosion” in the use of Webrecorder with 51% (=23) indicating its use, which is more than double the rating from the 2016 survey (Farrell et al. 2018, p. 20).

## 2.2 Web Archive User Studies

Costa and Silva (2010) conducted research for the Portuguese Web Archive (Arquivo.pt), to explore user intents and collect information on topics which are of most interest to users. Their method entails the collection of quantitative and qualitative data via 400 search logs, an online questionnaire (during the search process) (n=19), and a laboratory study (n=21). They found the majority of participants tended to use the full-text search, and had a preference for searching for older materials. For Costa and Silva (2010) this offers an indication that the value of a web archive increases as the web content gets older. Participants from the study suggest that it would be useful to view the evolution of a website/page over time or compare pages side-by-side. A personal space for a user to manage their search histories, and the ability to search for images is also mentioned. The top searched topics of the participants include computers/internet, education, health, commerce, and entertainment, with named individuals being the most searched topic.

## 2.3 Web Archiving Practises & Challenges for Using Web Archives

Sponsored by the Harvard Library, Truman (2016) conducted a study to document international web archiving programs (with a focus on cultural memory institutions), and examine the researcher use of web archives, and the barriers to working with web archives. Truman's methodology includes independent research and participation in working groups at conferences. It also entails semi-structured interviews or email communications with individuals from 23 institutions in the United States, Europe, and New Zealand with web archiving programs (or institutions intending to commence a programme), two service providers (n=2) and researchers who use web archives (n=4). Truman's (2016) study aims "to identify common concerns, needs, and expectations in the collection and provision of web archives to users; the provision and maintenance of web archiving infrastructure and services; and the use of web archives by researchers" (p. 6). From this, Truman (2016) notes that the main goal is "to identify opportunities for future collaborative exploration" (p. 6). In doing so, Truman examines how institutions provide and maintain their web archiving services and looks at the main challenges and gaps. How institutions integrate their web archives with their library collections, and others is also explored. Truman further provides a comprehensive directory of tools that have been developed to address the multiple functional needs across a lifecycle of web archiving, from selection, capture, and preservation, to access and tools used for research analysis. From the findings, Truman offers 22 opportunities for future research and development, organising them into four main themes as follows: increase communication and collaboration; focus on smart technical

development; focus on training and skills development; and build local capacity. While Truman suggests that the opportunities may fall under one or more themes, the number one theme is to increase collaboration and communication in several areas (Truman, 2016).

## 2.4 Web Archives and Scholarly Engagement

Costea (2018) conducted a study targeted at professors, researchers, and PhD students from the Arts, Humanities, and Social Sciences in two Danish universities, with the aims of providing some perspectives on scholarly engagement with web archives, reasons for non-use of web archives by researchers, and researcher needs in the use of web archives. Costea utilised a mixed method approach of an online survey (n=88), semi-structured interviews (n=3) and testing with first time users (n=2). Costea found there was a noteworthy lack of awareness of web archives as resources for research. Both users and non-users acknowledged the value of web archives, however, Costea suggests that to satisfy researchers' needs, there is a need for improvements to web archives in the areas of data selection, data management, discoverability options, and more access to methods for data analysis. Participants also mention issues related to the incompleteness of the data; thus, more comprehensive documentation and metadata is seen as a requirement for researchers. The findings also highlight a need for researchers to be able to extract data from a web archive to create a dataset for their own research requirements (Costea, 2018).

Healy (2021) conducted an online survey of lecturers, researchers, and students in Irish universities to gather information on the current state of scholarly awareness and engagement with web archives in Irish academic institutions. The survey consisted of questions to collect quantitative data, with a small element of qualitative data. The results of the survey are based on a final number of 239 respondents (N=239), of which 180 identified as non-users, and 59 participants identified as users. The findings suggest that the main reason for non-engagement with web archives in Irish academic institutions is due to a lack of awareness of the existence of web archives as resources for research. Other reasons for non-engagement are simply because web archives have no relevance for some research disciplines. Other reasons relate to challenges for using web archives such as search, navigation, and discovery functions, and dealing with large volumes of data. The representativeness of the data is presented as a challenge, in terms of understanding what gets preserved in a web archive, and what gets excluded. Other challenges include copyright implications for using archived web content, and how to provide a citation for sources from a web archive (Healy, 2021)

## 2.5 Research Data Management in Web Archive Studies

Truter (2021) offers one of the few studies which specifically looks at research data management and data sharing practices of researchers in 'Web Archive Studies'. Here Truter is referring to researchers who use web archives, and archived web data as part of their studies. Using a mixed methods approach, Truter's study combines a survey targeted at international Web Archive Studies researchers (n=31), and one semi-structured interview with an individual who has experience working with research data from web archives. For Truter, one of the main challenges for sharing archived web data/materials is legal restrictions, inclusive of copyright and third-party ownership, privacy policies, and GDPR, which creates challenges not only for the use of data from web archives but may also affect the ability to share the data or make it reusable. Truter's study further highlights challenges with the volume of data as well as the complexities of the data, with different media types and formats. The study participants also cite challenges such as a lack of a dedicated repository for the long-term preservation of archived web data; difficulty with Data Management Plans (DMPs); and a lack of storage space. Other challenges include a lack of funding for research data management, and a lack of guidance/training provided by publishers for those undertaking research in web archive studies (Truter, 2021).

# 3. METHODOLOGY

In this section, we lay out the methodological approach for the study, which includes the survey design, and approaches for data collection and analysis. The study was conducted in compliance with best practice guidelines for the collection and management of research data, as outlined in Maynooth University Research Ethics Policy (2020), Maynooth University Research Integrity Policy (2016, 2021), and Maynooth University Online Surveys User Policy (2019). The principal investigator acted as the data controller for the collection, storage, and preservation of the collected, and analysed data. Once the study is complete, the data will be prepared for migration to a location for long-term preservation on a private server repository in Maynooth University and will be preserved for a period of ten years, after which, it will be deleted in full (as outlined in MU Research Integrity Policy).

## 3.1 Survey Design and Questions

The survey was designed as an online questionnaire, to gather statistical and qualitative data in the form of free text responses. Our reasons for this method choice are based on factors such as cost and resource limitations due to it being a non-funded collaborative project. Also, Truter (2021) and the National Digital Stewardship Alliance (NDSA) have been successful in producing environmental data on web archive research with this type of model (NDSA Content Working Group, 2012; Bailey et al. 2014; Bailey et al. 2017; Farrell et al., 2018). Thus, we considered an online questionnaire to be a cost effective and relatively user-friendly method that would maximise responses.

Participants were not asked for any personal data such as Name/Contact Email/Date of Birth etc., and there were no IP addresses collected. However, participants were asked about their current country of residence, to observe the outreach of the survey, and to offer some insights on challenges which may be geographically relevant. While the data reveals some such connections, it was decided not to relate participants' responses to a particular geographical code. The web archive research community is a niche collaborative community, which tends to have a good knowledge of others in the field, therefore, we felt that using geographical codes may be problematic to retain anonymity. In addition, participants were asked about their age range and gender to explore whether age or gender has any relation to challenges to working with or using web archives. Participants were further asked about their positions and interests as a means to get an overall sense of the communities who work with and use the archived web. In compliance with good practice for collecting research data and

to minimise risks, participants were provided with information about the project, the time it would take to complete the questionnaire, an assurance of anonymity for responses, what the results would be used for, and contact information of the researchers involved. We also sought from participants their permission to publish extracts of text responses, to which most participants agreed. For those giving no permissions, their responses are aggregated into the coding system. Participants were also informed that they could withdraw at any time during the process of filling out the survey, and in doing so, their responses would not be collected.

The questionnaire was organised in 5 parts, and consisted of 28 questions, with a mix of tick box, multiple choice, Likert scales, and free text comment box answers. In Part 1, participants were asked to answer some demographic questions. In Part 2 participants were asked about the types of data they collect, their research outputs, the type of tools they use for data collection, and data analysis. Part 3 looked at the participants' skills and knowledge, while Part 4 examined citation systems, and challenges for citing archived web content. In part 5, participants were asked about the resources they found useful to further their skills and knowledge for working with/using web archives for research.

To test the navigation, and ensure the questions were clearly understood, the survey was pre-tested in mid-March 2021 by the research team, and 6 other colleagues from academic, non-academic, cultural heritage backgrounds. Nonetheless, a typing error was later discovered in the answer choices of one of the questions in the online survey (Q.16), when participation was already underway. We felt that the erroneous answer choices did not make sense in line with the question being asked, thus, it was decided not to include the responses from this section. However, a second part of the question provided participants with an ‘Other’ option, to enter free text, and is relative to the question being asked. Thus, it was decided to code this section, as a standalone result.

A final draft of the research project including information about the project, informed consent, a copy of the survey questions, and a data management plan were submitted to Maynooth University Research Ethics Committee, and the project received approval (SRESC-2021-2436150). A copy of the Information Sheet is attached as Appendix A, and a copy of the survey questions are attached as Appendix B.

## 3.2 Survey Software

The project utilised the JISC Online Surveys tool for collection purposes (Joint Information Systems Committee). Maynooth University provides access to this software for academic, and

research purposes to staff and PhD students. To note here, it is the only tool permitted by the university for conducting online survey studies of this nature.[5]

## 3.3 Survey Recruitment

The focus of the study is on individuals around the globe who participate in web archive research, in the context of web archiving, curation, and the use of web archives and archived web content for research or other purposes. However, we would like to point out that the global outreach of the web archiving community is limited. For example, Gomes, Miranda, and Costa (2011) provide an overview of global development in web archiving initiatives and observe that there was a significant growth in web archiving initiatives from 2003, but mostly in developed countries. Moreover, web archiving initiatives are more strongly represented in North America and Europe, as is evident from the 'List of Web archiving initiatives' (Wikipedia, 2011+).

The recruitment strategy consisted of recruitment emails to network lists for archivists, librarians, curators, digital humanities, internet studies, and web archive studies. The email also encouraged recipients to share amongst colleagues and networks. Examples of network lists include: AOIR members, IIPC curators and members, IFLA DIGLIB members, and WARCnet members. Recruitment also entailed social media posts for participation on Facebook, Twitter, and Slack, such as ADHO Facebook, EWA Twitter and the WARCnet Slack community.

## 3.4 Survey Responses

The survey was open from 21 July to 23 September 2021. We anticipated 25-30 complete questionnaires would be an acceptable level for the research. We based this in line with similar qualitative/quantitative studies such as Thomas et al. (2010) (n=17), Truman (2016) (n=23), and Truter (2022) (n=31). Overall, 50 participants responded to the survey. However, 6 surveys were removed from the survey dataset, due to some response inconsistencies. For example, some respondents seemed to confuse a web archive with other types of resources such as digital libraries, digital archives, or data repositories. In total, there were 6 such

[5] The use of the tool is subject to the terms and conditions set forth in Maynooth University Online Surveys User Policy, as well as Data Protection Laws (the GDPR and the Data Protection Act 2018), Maynooth University Responsible Computing Policy, and all applicable contracts and licences including Acceptable Statement Use issued by Online Surveys.

instances. Therefore, the final tally of complete surveys for analysis in this study is 44 respondents.

Other studies have also come across similar anomalies whereby there is some confusion with the term web archive (Costea, 2018; Healy, 2021). In a study conducted on awareness and engagement with web archives in Irish academic institutions, Healy (2021) found several instances of confusion whereby some respondents equated a web archive as being the same as a digital library, digital archive, or digital data repository. In a Danish study on scholarly awareness and engagement with web archives, Costea (2018, p. 11) also found some confusion with the term and suggests that the term web archive may not be "self-explanatory" enough for some researchers, and this could be due to "an ongoing lack of audience familiarity with the source." Indeed, Brügger (2018) also discusses the challenge with the term, but notes that while it may be confusing, the terms web archive and web archiving were coined decades ago and so, they are already part of the language for this resource type (pp. 77–78).

## 3.5 Survey Data Analysis

Some of the data was analysed through the JISC Online Surveys platform tools for filtering and aggregating data. Microsoft Excel was used for generating charts and graphs, which were exported as PNG files. The qualitative parts of the data were coded and analysed through MAXQDA (Release 20.3.0), a computer-assisted qualitative data analysis software (CAQDAS). While there are several commercial software available for coding qualitative data such as Atlas.ti or NVivo, and open source software such as Taguette or QualCoder, we utilised MAXQDA, as one of the research team members had access to a licence, and had experience using the software. The qualitative data analysis consisted of a process to examine and identify what the data represents, through a coding system of thematic representations. We further analysed the thematic representations (codes) through a critique of the codes, and a feedback-loop iterative process amongst the project team researchers. Also, to note here, several tables in the findings contain in-vivo representations. The term in-vivo comes from grounded theory and means that words or terms used by the respondents are so unique or insightful that they should be represented as standalone codes (MAXQDA Blog, 2021).

In relation to questions which contained free text responses for software and tools, we required desk research to assist in understanding the characteristics, and functionalities of the documented tools. To assist with this, we referred to the IIPC Tools & Software web page, and the NetLab Tools and Tutorials annotated directory. We also appealed to WARCnet members at the WARCnet Autumn 2021 hybrid meeting in Aarhus University, for assistance

in understanding the functionalities of some tools. In addition, we were hugely assisted by the addition of a research team member with a background in digital heritage and IT development, who showed great patience in explaining technical concepts to other members of the team.

## 3.6 Survey Limitations

Participation was voluntary, and participants could withdraw at any time during the process of filling out the survey, with the knowledge that their responses would not be collected. The questionnaire contained a mixture of both quantitative and qualitative answer options, taking an estimated 15 minutes to complete. This may have been off-putting and goes beyond the recommended time of 8-10 minutes which is generally used as a guideline to encourage completion (Chudoba, 2018; CoolTool, 2017; Steber, 2016). As mentioned previously in section 3.3, while the focus of the study is on individuals around the globe who participate in web archive research, the global outreach of the web archiving community is limited, and more strongly represented in North America and Europe. It is also worth noting that some professional fields are more represented in the data than others and is further discussed in section 4.1.2 Participant positions. Consequently, this may result in an over-representation of participants from some sectors. Nonetheless, we feel that this does not affect the overall aims of the research, in terms of developing an understanding of the current landscapes of web archive research. It is also worth noting, as with all studies based on survey sampling, this study cannot be construed to represent any target group population as a whole.

# 4. RESULTS & ANALYSIS

The results and analysis are based on a final number of respondents (N=44). Some percentages (%) and no. of participants (N/n=), are reflective of this, unless otherwise stated in the case of non-required questions. In addition, several sections are related to answers with free text responses, in these instances, the responses are analysed through the number of times a particular skill, tool, method, challenge etc. is mentioned in participants' answers. For instance, one participant may mention the use of a variety of tools for website capture, and each individual tool mentioned is included as a representation (R/r=).

## 4.1 Demographics

Overall, the respondents (N=44) identify with residing in North America, Europe, and Asia. This section further provides an overview of responses to questions on gender, position, and general research interests of the participants.

### 4.1.1 Participant age and gender

Provided with tick box options, participants were asked about their age range and gender.

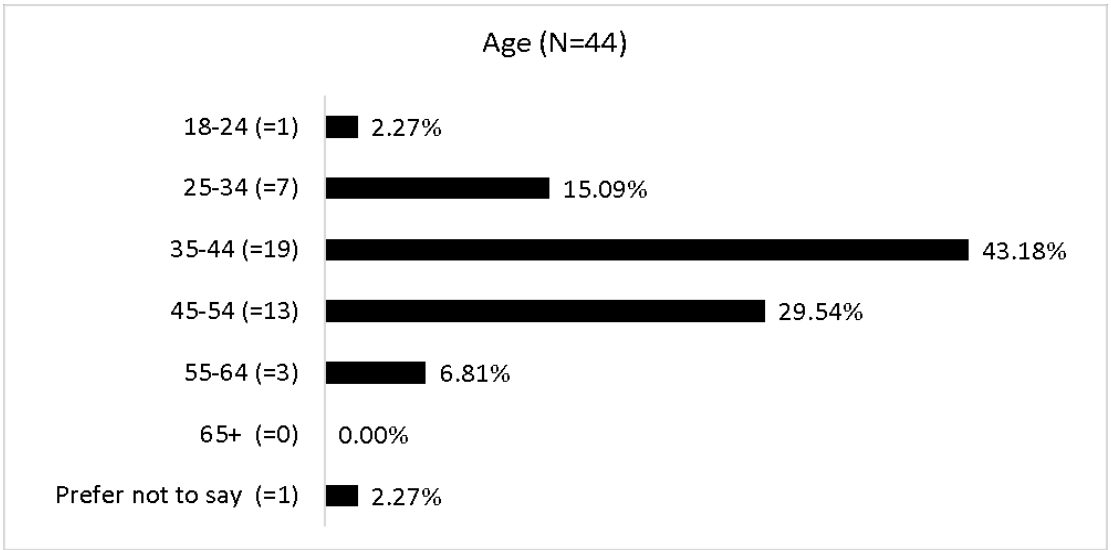


Figure 4.1: Representation of participant responses for age (N=44)

Figure 4.1 provides an overview of participant responses for age. Of overall participation (N=44), the highest representation age group is 35-44 years (43.18%, n=19), followed by the age groups of 45-54 years (29.54%, n=13), and 25-34 years (15.09%, n=7). Figure 4.2 provides

an overview of participant responses for gender (N=44) and shows an equal balance of female respondents (47.72%, n=21) and male respondents (47.72%, n=21).

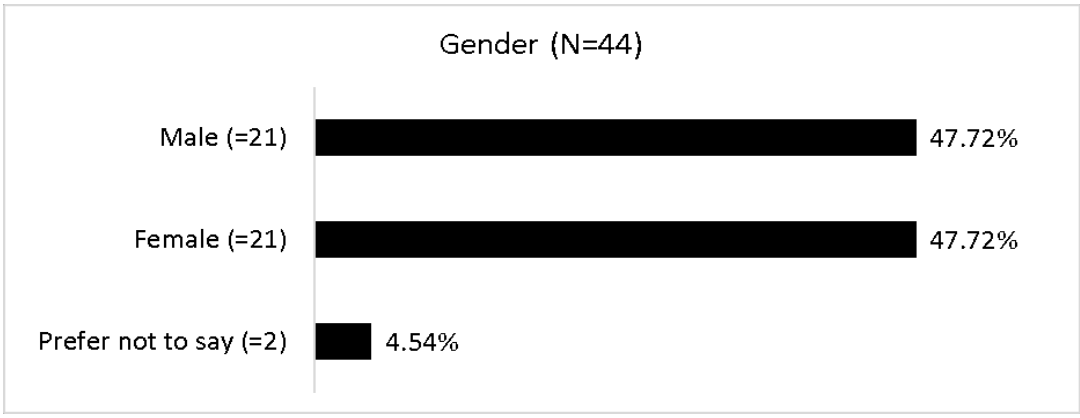


Figure 4.2: Representation of participant responses for gender (N=44)

### 4.1.2 Participant positions

Provided with a comment box, participants were asked to describe their position in their own words (e.g., PhD student in Media studies; Web archivist; IT specialist in a library; Senior lecturer in Sociology). All participants (N=44) provided free text which was coded into two main thematic representational categories.

As shown in Table 4.1, the first theme represents participants who identified with being employed in a Library, Archive, or Web Archive environment (n=30). To note, within this category, we also included respondents who identified with working in IT in a library/archive environment. The remaining participants (n=14) identified with being a scholar, academic, or lecturer, (n=9), a post-graduate/PhD student (n=2) or being employed in an IT or web design environment (n=3). Thus, we have labelled this group as Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=14). We acknowledge here that the individuals who identified with working in an IT or a web design environment, outside of academia, could have been categorised as a separate representation, but as they are such a small number, we included them in this categorical field, to minimise risks of identification through their responses.

Also, worth mentioning here, we initially thought it might be possible to align participants' positions with whether they were creators of web archives, or consumers/users of web archives, but this was not the case. For instance, some respondents in the Library, Archive, or Web Archive environment also indicate that they use other web archives as part of their workflows and research. Alternatively, some respondents in the Scholar, Academic, Lecturer,

Student, or IT/Web Design environment could also be considered as creators/curators of web archives for research purposes. Thus, the categorisation of participants' positions was not as clear-cut as originally imagined, and we acknowledge that there is some overlap.

Table 4.1: Thematic representation of participant responses for position (N=44)

| Theme representation for position (N=44) | Representation description | No. of participants |
|---|---|---|
| Library, Archive, or Web Archive environment | This refers to a participant who identifies with being employed in a Library, Archive, or Web Archive environment (including IT personnel). | n=30 |
| Scholar, Academic, Lecturer, Student, or IT/Web Design environment | This refers to a participant who identifies with being a Scholar, Academic, or Lecturer, a Postgraduate or PhD student; or a participant employed in an IT or Web Design environment. | n=14 |

### 4.1.3 Participant research interests in general

Participants were asked to describe their research interests in general in a comment box. All participants (N=44) provided free text responses which were coded into multiple thematic representations. 1 representation is in-vivo and offers another interpretation. The responses for this section are analysed through the number of times a particular research interest is mentioned and is documented as a representation (R/r=).

Table 4.2 offers an overview, and breakdown of such representations (N=44) which include the following:

- Information sciences (information studies) (r=55)
- Arts, Humanities, DH, Social Sciences, Media Studies (r=30)
- Internet/web applications, systems (r=7)
- IT/Computer applications, systems, environments (r=6)
- Research practises and approaches (r=5)
- Audiovisuals, Music, Video Games (r=4)
- Design related interests (r=4)
- Law (r=3)
- Transnationalism, Migration (r=2)
- Reading (r=1)

- Travel (r=1)
- In-vivo representations (r=1)

To note here, we use the theme 'Information sciences' (also known as information studies) in a broad sense. Wikipedia offers a useful description of information science as a "field which is primarily concerned with analysis, collection, classification, manipulation, storage, retrieval, movement, dissemination, and protection of information" (Wikipedia, 2002+). Within the theme of 'Information sciences' we include aspects of library and information sciences, archival science, museum studies, digital preservation, and forensics etc.

Table 4.2: Thematic representation of participant responses for their interests in general (N=44)

| Theme representation for participants' interests in general (N=44) | No. of representations (R=119) |
|---|---|
| \|> Information sciences (information studies)<br>● Web archives, web archiving, curation (=25)<br>○ Foster pathways for research access/use (r=14)<br>○ Collection development/strategies (r=4)<br>○ Web archiving/curation (in general) (=4)<br>○ Web archiving and metadata (r=2)<br>○ Web archives - compliancy for linked open data standards (r=1)<br>● Archives and records management (r=8)<br>● Digital preservation, long-term preservation (r=6)<br>● Libraries and digital libraries (r=7)<br>● Digital preservation, long-term preservation (=6)<br>● Documentation (institutional/organisational) (r=2)<br>● Media formats (r=2)<br>● Email archiving (r=1)<br>● Information literacy (r=1)<br>● Literature evolution (r=1)<br>● Museum studies (r=1)<br>● Open access and scholarly publication (r=1) | r=55 |
| \|> Arts, Humanities, DH, Social Sciences, Media Studies<br>● History (r=10)<br>● Culture and heritage (r=5)<br>● Languages, Linguistics, Semiotics (r=4)<br>● Identity and Memory (r=3)<br>● Anthropology (r=1)<br>● Archaeology (r=1) | r=30 |

| | |
|---|---|
| ● Cinema (r=1)<br>● Egyptology (r=1)<br>● Ethnography (r=1)<br>● Politics (r=1)<br>● Psychology (r=1)<br>● Sociology (r=1) | |
| \|> Internet/web applications, systems, histories<br>● Web design/ designers (r=2)<br>● Privacy and consent online (r=1)<br>● Vernacular web (r=1)<br>● Web based information systems (r=1)<br>● Web based learning (r=1)<br>● Web tracking (r=1) | r=7 |
| \|> IT/Computer applications, systems, environments<br>● User experience (UX) design (r=2)<br>● Artificial intelligence (r=1)<br>● Information technology (r=1)<br>● IT system architecture (r=1)<br>● Text recognition (r=1) | r=6 |
| \|> Research practises and approaches<br>● r: “archived web as a source”<br>● r: “evolving research practices with born digital material“<br>● r: “The impact of changing technology on historical research practice.”<br>● r: “Longitudinal in nature - both from a DH perspective and a technical one.”<br>● r: “digital methods for humanities research” | r=5 |
| \|> Audiovisuals, Music, Video Games | r=4 |
| \|> Design related interests<br>● Design & Anthropology (r=1)<br>● Design education (r=1)<br>● Design history (r=1)<br>● Design pedagogy (r=1) | r=4 |
| \|> Law<br>● Case law (r=1)<br>● Regulations (r=1)<br>● Legislation (r=1) | r=3 |

| | |
|---|---|
| \|> Transnationalism, Migration | r=2 |
| \|> Reading | r=1 |
| \|> Travel | r=1 |
| \|> In-vivo representations<br>● r: “Probably broader than they should be!“ | r=1 |

## 4.2 Data, Tools, and Methods

This section provides an overview of responses to questions on types of data collected, types of tools for data collection and analysis, and types of data outputs.

### 4.2.1 Types of data collected

Participants (N=44) were asked about the types of data they collect as part of their research in working with web archives and archived web content. Participants were offered several answer choices and an option of ‘Other’ to enter free text. Table 4.3 offers a breakdown of participant responses, in descending order of highest responses. A high number of respondents identified with collecting data such as URLS (68.88%, n=31); PDF files (64.44%, n=29) and WARC files (62.22%, n=28). This is followed by Archival metadata (55.55%, n=25), Images (53.33%, n=24), Screenshots (53.33%, n=24), Text files (51.11%, n=23), Numerical data (e.g., statistics) (44.44%, n=20), and Crawl logs (40.00%, n=18).

5 participants entered free text for other ‘Option’ as follows:

- Response: “social media content gathered via APIs”
- Response: “software”
- Response: “CDX index files, derivative crawl reports”
- Response: “Cascading Style Sheets, .json output from APIs, [...] JavaScript”
- Response: “tbc for outgoing work website”

Table 4.3: Breakdown of participant responses for the types of data they collect (N=44)

| Participant responses for the types of data they collect (N=44) | % of participants | No. of participants (N=44) |
|---|---|---|
| URLs | 68.88% | n=31 |
| PDF files | 64.44% | n=29 |
| WARC files | 62.22% | n=28 |
| Archival metadata | 55.55% | n=25 |
| Screenshots | 53.33% | n=24 |
| Images (e.g., photographs) | 53.33% | n=24 |
| Text files | 51.11% | n=23 |
| Numerical data (e.g., statistics) | 44.44% | n=20 |
| Crawl logs | 40.00% | n=18 |
| Audio files | 33.33% | n=15 |
| GIFs | 28.88% | n=13 |
| HTML code | 28.88% | n=13 |
| Banners | 20.00% | n=9 |
| Button Icons | 13.33% | n=6 |
| Tracking cookies | 13.33% | n=6 |
| 'Other' | 11.11% | n=5 |

### 4.2.2 Tools and methods for data collection

Provided with a comment box, participants were asked about the types of tools they use to 'Collect' their data. Of total participation (N=44), 41 participants provided free text comments which were coded into several thematic representations, and further bifurcated in line with the 2 thematic representations for participants positions as outlined in section 4.1.2. The responses for this section are analysed through the number of times certain tools or methods are mentioned and are documented as a representation (R/r=).

#### 4.2.2.1 Library, archive, or web archive environment

Table 4.4 offers a breakdown of the thematic representation for responses by participants who identified with working in a Library, Archive or Web Archive environment (n=30). 3 representations are in-vivo and offer other interpretations.

The thematic representations for tools and methods for data collection by these participants (n=30) include:

- Crawling software (r=37)
- Curating web archive collections: selection, configuring and scheduling crawls, annotating seeds, performing QA (r=10)
- Accessing/replaying archived web data (r=8)
- Managing data (r=5)
- Finding source material (r=4)
- Tools with diverse purposes (r=4)
- Collecting data from API (r=2)
- Screenshot, screen capture, screencast (r=2)
- Digital forensics/preservation (r=1)
- Web archiving subscription services (r=1)
- In-vivo representations (r=3)

Table 4.4: Thematic representation of responses for tools and methods used for data collection by participants who identified with Library, Archive, or Web Archive environment (n=30)

| Theme representation of responses for tools and methods used for data collection by participants who identified with Library, Archive, or Web Archive environment (n=30) | No. of representations (R=77) |
|---|---|
| \|> Crawling software<br>● Browser-based crawlers (r=23)<br>○ Conifer (prior, Webrecorder) (r=9)<br>○ ArchiveWeb.page (r=4)<br>○ Brozzler (r=4)<br>○ Electrolyte (r=3)<br>○ Browsertrix (r=2)<br>○ Umbra (r=1)<br>● Crawl software in general, not browser-based (r=13)<br>○ Heritrix (r=11)<br>○ HTTrack Website Copier (r=1)<br>○ Wget (r=1) | r=37 |

| | |
|---|---|
| ● Web crawler (in general) (r=1) | |
| \|> Curating web archive collections: selection, configuring and scheduling crawls, annotating seeds, performing QA<br>● NetarchiveSuite (r=5)<br>● CWeb (r=2)<br>● W3ACT (r=1)<br>● Web Curator Tool (r=1)<br>● r: "selecting material for collection" | r=10 |
| \|> Accessing/replaying archived web data<br>● Internet Archive, Wayback machine (r=3)<br>● OpenWayback (r=2)<br>● pywb (r=2)<br>● waybackpy (r=1) | r=8 |
| \|> Managing data<br>● Excel, spreadsheet, .csv (r=3)<br>● CMS, Cloud platforms (r=2)<br>○ DSpace (r=1)<br>○ Google Drive (r=1) | r=5 |
| \|> Finding source material (r=4)<br>● Internet, search engines, web search (r=2)<br>● Library catalogues and databases (r=2) | r=4 |
| \|> Tools with diverse purposes (=4)<br>● Browser tools (r=1)<br>● command-line tools (r=1)<br>● Python scripts/libraries (r=1)<br>● r: "the type of tools that come for standard with a PC" | r=4 |
| \|> Collecting data from API<br>● Instaloader (r=1)<br>● Social Feed Manager (r=1) | r=2 |
| \|> Screenshot, screen capture<br>● screen capture tools (in general) (r=1)<br>● snipping tools (in general) (r=1) | r=2 |
| \|> Digital forensics/preservation<br>● MediaArea tools (r=1) | r=1 |
| \|> Web archiving subscription services<br>● Archive-It (r=1) | r=1 |

| | |
|---|---|
| \|> In-vivo representations<br>● r: "In house developed web archiving tools""<br>● r: "institutional sources"<br>● r: " text recognition evaluation tools" | r=3 |

### 4.2.2.2 Scholar, academic, lecturer, student, or IT/web design environment

Table 4.5 provides a thematic representation of responses by participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=11). 3 representations are in-vivo and offer other interpretations.

The thematic representations for the tools and methods for data collection of these participants (n=11) include:

- Crawling software (r=7)
- Finding source material (r=6)
- Screenshot, screen capture, screencast (r=5)
- Tools with diverse purposes (r=4)
- File downloads (r=3)
- Accessing/replaying archived web data (r=2)
- Collecting data from API (r=2)
- Managing data (r=2)
- Web scraping (extracting data from web pages) (r=2)
- Audio tools (r=1)
- Curating web archive collections: selection, configuring and scheduling crawls, annotating seeds, performing QA (r=1)
- Manual collection for close reading (r=1)
- Web archiving subscription services (r=1)
- In-vivo representations (r=3)

Table 4.5: Thematic representation of responses for tools and methods used for data collection by participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=11)

| Theme representation of responses for tools and methods used for data collection by participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=11) | No. of representations (R=40) |
|---|---|
| \|> Crawling software<br>● Browser-based crawlers (r=3)<br>○ Conifer (prior, Webrecorder) (r=2)<br>○ Browsertrix (r=1)<br>● Crawl software in general, not browser-based (r=4)<br>○ Heritrix (r=2)<br>○ HTTrack Website Copier (r=1)<br>○ Wget (r=1) | r=7 |
| \|> Finding source material<br>● In libraries/web archives (r=3)<br>○ SHINE tools - UKWA (r=2)<br>○ Library catalogues and databases (r=1)<br>● Internet, search engines, web search (r=3)<br>○ Internet (r=1)<br>○ Search engines / web search (r=2) | r=6 |
| \|> Screenshot, screen capture, screencast<br>● screenshot tools/functions (in general) (r=2)<br>● script for screenshot automation (r=1)<br>● Snagit (r=1)<br>● Websnapper (r=1) | r=5 |
| \|> Tools with diverse purposes<br>● Browser tools (r=2)<br>● Python scripts/libraries (r=1)<br>● R (Rstudio) (r=1) | r=4 |
| \|> Manual/scripted file downloads<br>● save files manually (r=1)<br>● manual/scripted downloads (r=1)<br>● general file download (r=1) | r=3 |
| \|> Accessing/replaying archived web data<br>● Internet Archive, Wayback machine (r=2) | r=2 |

| | |
|---|---|
| \|> Collecting data from API<br>● Twarc (=1)<br>● r: "make my own tools to collect data based on [publicly] available API" | r=2 |
| \|> Managing data<br>● Citation and reference management (r=2)<br>○ Zotero (r=1)<br>○ Zotfile PlugIn (r=1) | r=2 |
| \|> Web scraping (extracting data from web pages)<br>● Webscraper.io (=1)<br>● web scraping scripts (=1) | r=2 |
| \|> Audio tools (for interviews)<br>● r: "audio recording tools (for interviews), etc." | r=1 |
| \|> Curating web archive collections: selection, configuring and scheduling crawls, annotating seeds, performing QA<br>● Web Archiving Integration Layer (WAIL) (r=1) | r=1 |
| \|> Manual collection for close reading<br>● r: "I mostly do it [manually], as I work with close reading" | r=1 |
| \|> Web archiving subscription services<br>● Archive-It (r=1) | r=1 |
| \|> In-vivo representations<br>● r: "non-English language search words"<br>● r: "direct contact with people who might have the data"<br>● r: "scanning/OCR if the source is hard copy" | r=3 |

### 4.2.3 Tools and methods for data analysis

Provided with a comment box, participants were asked about the types of tools and methods they use to 'Analyse' their data. Of total participation (N=44), 36 participants provided free text comments which were coded into several thematic representations, and further bifurcated in line with the 2 thematic categories for participants positions as outlined in section 4.1.2. The responses for this section are analysed through the number of times a particular tool or method is mentioned and is documented as a representation (R/r=).

#### 4.2.3.1 Library, archive, or web archive environment

Table 4.6 offers a breakdown of the thematic representation for responses by participants who identified with working in a Library, Archive or Web Archive environment (n=25). 3 representations are in-vivo and offer other interpretations.

The thematic representations for tools and methods for data collection by these participants (n=25) include:

- Search and information retrieval (r=13)
- Data extraction, cleaning, transformation (r=6)
- Programming/scripting languages, computing environments (r=6)
- Visualisation (r=4)
- Digital forensics/preservation (r=3)
- Distributed processing (r=3)
- Metadata, crawl logs (r=3)
- Network analysis (r=3)
- Replay/playback tools (r=2)
- Computer-assisted text analysis (r=2)
- Data management (r=2)
- Collaboration (r=1)
- Computing infrastructure (r=1)
- Evidence analysis (r=1)
- Machine learning (r=1)
- Statistics (in general) (r=1)
- Web archive access and analysis (r=1)
- Web archiving management (r=1)
- In-vivo representations (r=3)

Table 4.6: Thematic representation of responses for tools and methods used for data analysis by participants who identified with Library, Archive, or Web Archive environment (n=25)

| Theme representation of responses for tools and methods used for data analysis by participants who identified with Library, Archive, or Web Archive environment (n=25) | No. of representations (R=58) |
|---|---|
| \|> Search and information retrieval<br>● CDX queries/files (r=2)<br>● SolrWayback (r=2)<br>● SQL (r=2)<br>● Amazon Athena (AWS) (r=1)<br>● Apache Solr (r=1)<br>● ElasticSearch (r=1)<br>● HeidiSQL/MariaDB (r=1)<br>● Apache Lucene (r=1)<br>● NutchWax (r=1)<br>● r: "Web Archive user interface, faceted functions" | r=13 |
| \|> Data extraction, cleaning, transformation<br>● Excel, spreadsheets (r=5)<br>● Archives Unleashed Toolkit (r=1) | r=6 |
| \|> Programming/scripting languages, computing environments<br>● Python/Python libraries (r=3)<br>● Command-line tools (r=1)<br>● Jupyter Notebooks (r=1)<br>● R (r=1) | r=6 |
| \|> Visualisation<br>● Tableau (r=2)<br>● Kibana (r=2) | r=4 |
| \|> Digital forensics/preservation<br>● DROID (r=1)<br>● BitCurator (r=1)<br>● MediaArea tools (r=1) | r=3 |
| \|> Distributed processing<br>● Apache Hadoop (r=2)<br>● Apache Spark (r=1) | r=3 |
| \|> Metadata, crawl logs<br>● Crawl logs (r=2) | r=3 |

| | |
|---|---|
| ● r: "Metadata" | |
| \|> Network analysis<br>● Gephi (r=3) | r=3 |
| \|> Replay/playback tools<br>● OpenWayback (r=1)<br>● Pywb (r=1) | r=2 |
| \|> Computer-assisted text analysis<br>● IramuteQ (r=1)<br>● Voyant tools (r=1) | r=2 |
| \|> Data management<br>● Apache Parquet (r=1)<br>● Excel, spreadsheets (r=1) | r=2 |
| \|> Collaboration<br>● r: "brainstorming with colleagues" | r=1 |
| \|> Computing infrastructure<br>● Amazon Web Services (r=1) | r=1 |
| \|> Evidence analysis<br>● r: "I collect it for lawyers who analyze it." | r=1 |
| \|> Machine learning<br>● TensorFlow (r=1) | r=1 |
| \|> Statistics (in general) (=1) | r=1 |
| \|> Web archive access and analysis<br>● GLAM workbench notebooks (r=1) | r=1 |
| \|> Web archiving management<br>● Digiboard (r=1) | r=1 |
| \|> In-vivo representations<br>● r: "lists, notes, tiny pieces of paper"<br>● r: "manual statistics on the report files [from SolrWayback]"<br>● r: "My work with the web archive involves selecting material, not carrying out research." | r=3 |

### 4.2.3.2 Scholar, academic, lecturer, student, or IT/web design environment

Table 4.7 provides a thematic representation of responses by participants who identified with being a Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=13). 2 representations are in-vivo and offer other interpretations.

The thematic representations for the tools and methods for data collection of these participants (n=13) include:

- Data analysis, extraction, cleaning, transformation (r=8)
- Programming, scripting languages and computing environments (r=8)
- Qualitative data analysis (r=6)
- Network analysis (r=3)
- Other Tools (r=3)
- Collaboration (r=1)
- Computer-assisted text analysis (r=1)
- Visualisation (r=1)
- In-vivo representations (r=2)

Table 4.7: Thematic representation of participant responses for tools and methods used for data analysis by participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=13)

| Theme representation of responses for tools and methods used for data analysis by participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=13) | No. of representations (R=34) |
|---|---|
| \|> Data analysis, extraction, cleaning, transformation<br>● Excel, spreadsheets (r=4)<br>● Archives Unleashed Cloud (r=1)<br>● Archives Unleashed Toolkit (r=1)<br>● OpenRefine (r=1)<br>● Pattern matching (r=1)<br>● Regular expressions (r=1) | r=8 |
| \|> Programming, scripting languages and computing environments<br>● Bash/shell scripting languages (r=3)<br>● Python/Python libraries (r=2)<br>● Command-line tools (r=1)<br>● Perl (r=1)<br>● R (r=1) | r=8 |

| | |
|---|---|
| \|> Qualitative data analysis<br>● Nvivo (r=2)<br>● Atlas.ti (r=1)<br>● r: "annotating PDFs with PDFExpert"<br>● r: "Close reading of websites and it's html code"<br>● r: "manual qualitative content analysis" | r=6 |
| \|> Network analysis<br>● Gephi (r=3) | r=3 |
| \|> Other tools<br>● Microsoft 365 (r=1)<br>● Proprietary tools (r=1)<br>● r: "I usually make my own tools" | r=3 |
| \|> Collaboration<br>● Confluence (r=1) | r=1 |
| \|> Computer-assisted text analysis<br>● Voyant tools (r=1) | r=1 |
| \|> Visualisation<br>● r: "visualisation tools for qualitative data" | r=1 |
| \|> In-vivo representations<br>● r: "mostly my brain"<br>● r: "Conceptual tools (e.g. social semiotics, multimodality) for the [analysis] of complex web objects" | r=2 |

### 4.2.4 Types of data outputs

Provided with a comment box, participants were asked to describe the types of data they 'Output' as part of their research in working with web archives. Of total participation (N=44), 37 participants provided free-text responses which were coded into several thematic representations. 3 representations are in-vivo and offer other interpretations. The responses for this section are analysed through the number of times a particular type of data is mentioned and is documented as a representation (R/r=).

Table 4.8 offers an overview of the thematic representations which include:

- Excel, spreadsheets, .csv files (r=19)
- Screenshots (r=13)
- Text related (r=11)

- Visualisations (r=10)
- Web related, protocols, mark-up languages (r=7)
- Images/ Image collections (r=5)
- Metadata, crawl logs, indexes (r=4)
- Tables (r=4)
- Annotations, information summaries (r=3)
- Meta mark-up languages (r=3)
- Papers, articles, guides (r=3)
- PDF files (r=3)
- Collection development/selection (r=2)
- Multi-media outputs (r=2)
- Statistics (r=2)
- APIs (r=1)
- Digital forensics/preservation (r=1)
- Evidence collection (r=1)
- WARC files (r=1)
- In-vivo representations (r=3)

Table 4.8: Thematic representation of participant responses for types of data they 'Output' as part of their research in working with web archives (n=37)

| Theme representation for types of data outputs (n=37) | No. of representations (R=98) |
|---|---|
| \|> Excel, spreadsheets, .csv files<br>● Spreadsheets (r=16)<br>● .csv files (r=2)<br>● Excel (r=1) | r=19 |
| \|> Screenshots | r=13 |
| \|> Text related<br>● Text fragments/extracts (r=7)<br>● Quotes (r=2)<br>● Text (r=2) | r=11 |
| \|> Visualisations<br>● Graphs (r=5)<br>● Charts (r=2)<br>● Diagrams (r=1) | r=10 |

| | |
|---|---|
| ● Visualisations (in general) (r=1)<br>● Gephi, network analysis visuals (r=1) | |
| \|> Web related, protocols, mark-up language<br>● Web pages (r=2)<br>● HTML (r=1)<br>● Reconstructed web pages (r=1)<br>● Websites (r=1)<br>● Web statistics (r=1)<br>● URLs (r=1)<br>● r: "List of in- and outgoing links" | r=7 |
| \|> Images/ Image collections<br>● Images (r=2)<br>● image collections (r=1)<br>● image fragments (r=1)<br>● JPG (r=1) | r=5 |
| \|> Metadata, crawl logs, indexes<br>● Crawl logs (r=1)<br>● Metadata (r=2)<br>● Indexes (r=1) | r=4 |
| \|> Tables | r=4 |
| \|> Annotations, information summaries<br>● r: "Annotation summaries"<br>● r: "bulleted lists of findings"<br>● r: "summaries of information" | r=3 |
| \|> Meta markup languages<br>● XML (r=2)<br>● JSON (r=1) | r=3 |
| \|> Papers, articles, guides<br>● Papers written in LaTeX (r=1)<br>● Papers related to event collection (r=1)<br>● Research guides (r=1) | r=3 |
| \|> PDF files | r=3 |
| \|> Collection development/selection<br>● r: "selecting material"<br>● r: "special collection" | r=2 |

| | |
|---|---|
| \|> Multi-media outputs<br>● Twitter tweets (r=1)<br>● Wiki content (r=1) | r=2 |
| \|> Statistics | r=2 |
| \|> APIs | r=1 |
| \|> Digital forensics/preservation<br>● r: "Reports from BitCurator" | r=1 |
| \|> Evidence collection<br>● r: "The lawyers who I send it to publish research and use it in court cases." | r=1 |
| \|> WARC files | r=1 |
| \|> In-vivo representations<br>● r: "Image/textual search services online"<br>● r: "structured corpora”<br>● r: "I don't generate data myself. I would like to work more with visualisation and interpretation tools (eg Dark and Stormy archives project)" | r=4 |

## 4.3 Skills and Knowledge

This section looks at participants' primary areas of research with web archives, their reasons for curating/using web archives, the length of time working with web archives, the type of web archive services they use, and the types of challenges they encountered when curating/using web archives.

### 4.3.1 Primary areas of research/curation with web archives

Provided with a comment box, participants were asked to describe, in their own words, their primary areas of research/curation with web archives. All participants (N=44) provided free text, which was coded into several thematic representations. As mentioned earlier in section 4.1.3, we use the theme information science (also known as information studies) in a broad sense, and include aspects of library and information science, archival science, museum studies, digital preservation, and forensics etc., within this theme.

Table 4.9 offers an overview and breakdown of the thematic representation which include:

- Information sciences (information studies) (r=38)
- Arts, Humanities, DH, Social Sciences, Media Studies (r=23)
- IT, Computer, Web applications, systems (r=9)
- Audiovisuals, Music, Video Games (r=4)
- Politics (r=2)
- Business need (r=2)

Table 4.9: Thematic representation of participant responses for primary areas of research/curation with web archives (N=44)

| Theme representations for primary areas of research/curation with web archives (N=44) | No. of representations (R=78) |
|---|---|
| \|> Information sciences (information studies)<br>● Web archives, web archiving, curation (r=29)<br>○ Collection development (r=5)<br>○ Crawling (r=3)<br>○ Preservation (r=3)<br>○ Quality assurance (r=3)<br>○ Web archiving (in general) (r=3)<br>○ Curatorial management (r=2)<br>○ Promoting use of web archives for research (r=3)<br>○ Comparing transnational collection/curatorial processes (r=1)<br>○ Curating web archive collections for research (r=1)<br>○ Evaluating archival rate of national websites (r=1)<br>○ Information retrieval (r=1)<br>○ Metadata (r=1)<br>○ Social media archiving (r=1)<br>○ Web archive solutions (r=1)<br>○ Web archiving, history/evolution (r=1)<br>● Documentation & publications (r=5)<br>● Archival studies (r=2)<br>● Libraries and social media communications | r=38 |
| \|> Arts, Humanities, DH, Social Sciences, Media Studies<br>● Internet and web histories (r=7)<br>○ r: "internet literature history"<br>○ r: "Historical studies of the development of the [...] web"<br>○ r: "History of the [national] internet" | r=23 |

| | |
|---|---|
| ○ r: "history of websites (and the user experience of that) at the web archives"<br>○ r: "what kind of educational application there were on the web"<br>○ r: "web history"<br>○ r: "vernacular creativity on the [...] web"<br>● History (=4)<br>● Culture and heritage (r=3)<br>● Media related studies (r=3)<br>○ TV (r=1)<br>○ Media practises (r=1)<br>○ r: "I use web archives to track down information, particularly news stories and press releases, that is no longer available on any website"<br>● Antiquarian materials (r=1)<br>● Diasporic research (r=1)<br>● Education (r=1)<br>● Egyptology (=1)<br>● Ethnography (r=1)<br>○ r: "immersive methodologies (ethnography)"<br>● Online religion (r=1) | |
| \|> IT, Computer, Web applications, systems<br>● Evolution of the web (r=1)<br>● HTML Code (r=1)<br>● Influence of other forms of design on web design (r=1)<br>● Internet measurements (r=1)<br>● Link structures of the web (r=1)<br>● Responsive web design techniques (r=1)<br>● Web design and designers (r=1)<br>● Web design communities, and best practices (r=1)<br>● Web tracking techniques (r=1) | r=9 |
| \|> Audiovisuals, Music, Video Games | r=4 |
| \|> Business case<br>● Web content strategy<br>○ r: "My team uses web archives to understand how we presented content to customers in the past, to inform our current content strategies and experience design iteration plans"<br>● Collecting evidence for a law firm<br>○ r: "I collect it for lawyers who analyze it" | r=2 |

| | |
|---|---|
| \|> Politics | r=2 |

## 4.3.2 Reasons which led to curating/using web archives

Provided with a comment box, participants were asked about the reasons which led them to using web archives for their research. 42 participants provided free text responses which were coded into multiple thematic representations, and further organised in line with the 2 thematic categories for participants positions as outlined in section 4.1.2. The responses for this section are analysed through the number of times a particular reason is mentioned and is documented as a representation (R/r=).

### 4.3.2.1 Library, archive, or web archive environment

Table 4.10 offers a breakdown of the thematic representation for responses by participants who identified with working in a Library, Archive or Web Archive environment (n=28). 4 representations are in-vivo and offer other interpretations.

The thematic representations for the reasons which led these participants (n=28) to curating/using web archives include:

- Web archives, web archiving, curation (r=23)
- Concerns about the loss/changes of web content (r=3)
- Interests in research aspects/outputs of collections (r=2)
- Resource to find information/literature (r=2)
- Business need for a law firm library (r=1)
- Digital collection/curation (r=1)
- Library internship (r=1)
- Subject librarianship (r=1)
- In-vivo representations (r=4)

Table 4.10: Thematic representation of responses for reasons which led to curating/using web archives, by participants who identified with Library, Archive, or Web Archive environment (n=28)

| Theme representation of reasons which led to curating/using web archives, by participants who identified with Library, Archive, or Web Archive environment (n=28) | No. of representations (R=38) |
|---|---|
| \|> Web archives, web archiving, curation<br>● Web archivist/curator - job related (r=11)<br>● Promote/support research engagement with web archives (r=4)<br>● Institutional need (r=2)<br>● Digital legal deposit (r=1)<br>● Promote inclusive archiving (r=1)<br>● Promote value of web archives to stakeholders/funders (r=1)<br>● r: "It is the present and future of archival work."<br>● r: "A specific collection for a current [...] senator requires capturing his current website"<br>● r: "The later development of archival tools to capture and catalog websites has been invaluable" | r=23 |
| \|> Concerns about the loss/changes of web content<br>● Preserve documentary heritage (r=1)<br>● r: "As the field of archival science has developed, my interest has turned toward the mountain of data being produced and changed on the internet."<br>● r: "Loss of content as websites/databases are updated/retired/allowed to fail" | r=3 |
| \|> Interests in research aspects/outputs of collections<br>● r: “as a librarian I would like to work with the research aspect of this broad topic not just taking an overview from the curatorial perspective.”<br>● r: “I have degrees from History and European Studies, so I am interested in the various kind of research outputs of the collection.” | r=2 |
| \|> Resource to find information/old websites<br>● r: "I found it was easier to track down certain bits of information via web archives than it was to ask the organization for a past press release."<br>● r: "old websites as primary sources from about a decade ago" | r=2 |
| \|> Business need for a law firm library<br>● r: "It was the only source that had the information I needed" | r=1 |

| | |
|---|---|
| \|> Digital collection/curation | r=1 |
| \|> Library internship | r=1 |
| \|> Subject librarianship | r=1 |
| \|> In-vivo representations<br>● r: "Availability during pandemic"<br>● r: "An adviser taught me how to use it."<br>● r: "Internet Archive's Wayback Machine was an early fascination of mine."<br>● r: "My PhD Thesis" | r=4 |

#### 4.3.2.2 Scholar, academic, lecturer, student, or IT/web design environment

Table 4.11 provides a thematic representation of responses by participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=14). 3 representations are in-vivo and offer other interpretations.

The thematic representations for the reasons which led these participants (n=14) to curating/using web archives include:

- Resource for conducting research (r=10)
- Concerns about the loss of web content (r=2)
- Ease of access to public web archives (r=2)
- Resource to find information/old websites (r=2)
- Business need for web content strategy (r=1)
- Richness of data (r=1)
- In-vivo representations (r=3)

Table 4.11: Thematic representation of responses for reasons which led to using web archives for research, by participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=14)

| Theme representation of reasons which led participants to using web archives for their research, by participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=14) | No. of representations (R=21) |
|---|---|
| \|> Resource for conducting research<br>● Resource for historical research (r=3)<br>● Resource for studying migrants/migration (r=2)<br>● Resource for research of evolution of web design (r=1)<br>● Resource for research of educational broadcasting (r=1)<br>● Resource for internet studies research (r=1)<br>● r: "authoritative source" for research<br>● r: "The power of 'raw' internet data to triangulate other data and therefore add to the overall 'scientific' objectivity and credibility of the research" | r=10 |
| \|> Concerns about the loss of web content<br>● Website obsolescence (r=1)<br>● Preservation for the future (r=1) | r=2 |
| \|> Ease of access to public web archives<br>● r: "Having ready access to web archives, which coincided with emerging research questions"<br>● r: "Ease of access" | r=2 |
| \|> Resource to find information/literature<br>● r: "Wanting to find data"<br>● r: "online literary magazine which is not live again but important evidences in [...] literary history" | r=2 |
| \|> Business need<br>● Web content strategy<br>○ r: "My team uses web archives to understand how we presented content to customers in the past, to inform our current content strategies and experience design iteration plans" | r=1 |
| \|> Richness of data | r=1 |
| \|> In-vivo representations<br>● r: "Fascination with the centrality of the web in everyday lives and yet its propensity to obsolescence and research oversight" | r=3 |

| | |
|---|---|
| • r: "Wanting [to] make data available"<br>• r: "Web archiving is [a] very important topic, which is not researched enough" | |

### 4.3.3 Length of time curating/using web archives

Provided with multiple choice options, and time ranges, participants were asked about the length of time they had been using web archives for their research. Figure 4.3 provides an overview for respondents' answers (N=44). From this we can surmise that respondents are at novice, intermediate and experienced levels within web archive research.

Participant responses (N=44) indicates the following:

- 0-6 months (n=2)
- 6 months - 1 year (6.81%, n=3)
- 1-2 years (22.72%, n=10)
- 3-5years (15.90%, n=7)
- 5-10 years (25.00%, n=11)
- 10-15 years (15.90%, n=7)
- More than 15 years (9.09%, n=4)

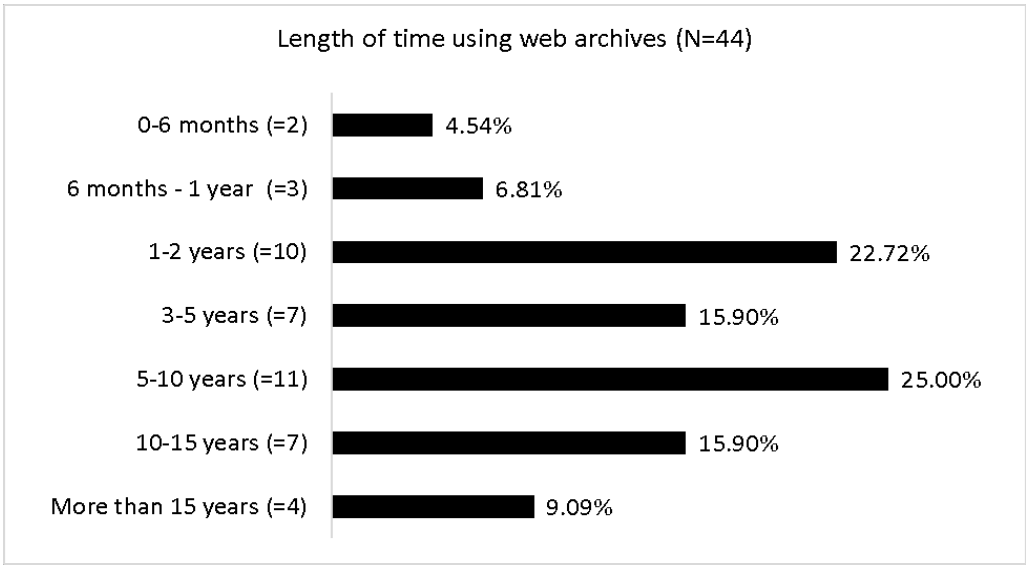


Figure 4.3: Representation of participant responses for the length of time using web archives (N=44)

### 4.3.4 Web archive providers and services

Participants were asked about the web archive(s) or services they use for their research, and offered several answer choices, and the option of 'Other' to enter free text.

Table 4.12 provides a full breakdown of responses, and we highlight some of the responses below in order of highest representation n ≥ 3. The full list of providers and services is provided as in the Bibliography.

- Wayback Machine (Internet Archive) (81%, n=36)
- UK Web Archive (British Library/UK Legal Deposit Libraries) (36.36%, n=16)
- Memento Time Travel (25.00%, n=11)
- US Library of Congress Web Archive (29.54%, n=13)
- UK Government Web Archive (The National Archives, UK) (22.72%, n=10)
- Arquivo.pt (FCT | FCCN, Portugal) (18.18%, n=8)
- Netarkivet (Royal Library, and the State and University Library, Denmark) (15.90%, n=7)
- Common Crawl (11.36%, n=5)
- UK Parliament Web Archive (UK Parliamentary Archives) (11.36%, n=5)
- BnF Archives de l'internet (Bibliothèque nationale de France) (9.09%, n=4)
- Archive.today (6.81%, n=3)
- INA Web Archive (Institut Nationale de l'Audiovisuel) (6.81%, n=3)
- Webarchief van Nederland (Koninklijke Bibliotheek) (6.81%, n=3)

Further to this, participants (n=14) provided free text for the 'Other' option. The free text was coded into several thematic representations.

Table 4.13 provides an overview of such representations which includes:

- Archivo de la Web Española (Biblioteca Nacional de España) (r=3)
- National Széchényi Library Web Archive, Hungary (r=2)
- Archive-It Collections (r=1)
- Archives Unleashed (r=1)
- Conifer (prior, Webrecorder) (r=1)
- Croatian Web Archive (HAW) (r=1)
- GLAM Workbench (r=1)
- International Internet Preservation Consortium (r=1)
- JISC UK web archive (1996-2013) / SHINE (r=1)

- National Records of Scotland Web Archive (r=1)
- Oldweb.today (r=1)
- Personal archives of early webmasters (r=1)
- WARC files created by a research project (r=1)

Table 4.12: Representation of participant responses for the web archive(s) or services they use (N=44)

| Answer Choices for web archive(s) or services used (N=44) | No. of participants | |
|---|---|---|
| Wayback Machine (Internet Archive) | 81.81% | n=36 |
| UK Web Archive (British Library/UK Legal Deposit Libraries) | 36.36% | n=16 |
| US Library of Congress Web Archive | 29.54% | n=13 |
| Memento Time Travel | 25.00% | n=11 |
| UK Government Web Archive (UK National Archives) | 22.72% | n=10 |
| Arquivo.pt (FCT \| FCCN, Portugal) | 18.18% | n=8 |
| Netarkivet (Danish Royal Library, and the State and University Library) | 15.90% | n=7 |
| Common Crawl | 11.36% | n=5 |
| UK Parliament Web Archive (UK Parliamentary Archives) | 11.36% | n=5 |
| BnF Archives de l'internet (Bibliothèque nationale de France) | 9.09% | n=4 |
| Archive.today | 6.81% | n=3 |
| INA Web Archive (Institut Nationale de l'Audiovisuel) | 6.81% | n=3 |
| Webarchief van Nederland (Koninklijke Bibliotheek) | 6.81% | n=3 |
| Luxembourg Web Archive (Bibliothèque Nationale de Luxembourg) | 4.54% | n=2 |
| Government of Canada Web Archive (Library and Archives Canada) | 2.27% | n=1 |
| NLI Web Archive (National Library of Ireland) | 2.27% | n=1 |
| PRONI Web Archive (Public Records Office of Northern Ireland) | 2.27% | n=1 |
| Other representations: | 34.09% | n=15 |

Table 4.13: Thematic representations of participant responses for 'Other' web archive(s) or services used (n=14)

| Theme representations for 'Other' web archives/services used (n=14) | No. of representations (R=18) |
|---|---|
| Archivo de la Web Española (Biblioteca Nacional de España) | r=3 |
| National Széchényi Library Web Archive, Hungary | r=2 |
| Archive-It Collections | r=1 |
| Archives Unleashed | r=1 |
| archives.design | r=1 |
| Conifer | r=1 |
| Croatian Web Archive (HAW) | r=1 |
| General State Archives of Greece | r=1 |
| GLAM Workbench | r=1 |
| International Internet Preservation Consortium | r=1 |
| JISC UK web archive (1996-2013) on the SHINE interface | r=1 |
| National Records of Scotland Web Archive | r=1 |
| Oldweb.today | r=1 |
| Personal archives of early webmasters | r=1 |
| WARC files created by a research project | r=1 |

### 4.3.5 Challenges encountered when working with web archives

Provided with a comment box, participants were asked to describe the challenges they encountered when working with web archives and discuss any workarounds. 41 participants provided free text which was coded into multiple thematic representations. It was also further organised in line with the 2 categories for participants positions as outlined in section 4.1.2. The responses are analysed through the number of times a particular challenge is

mentioned throughout the responses for this section and is documented as a representation (R/r=).

#### 4.3.5.1 Library, archive, or web archive environment

In relation to challenges, and participants who identified with working in a Library, Archive, or Web Archive environment, 27 participants provided free text responses. 2 participants specified that they encountered no challenges when working with web archives.

Table 4.14 offers an overview and breakdown of representations for the remaining participants (n=25).

Representations for challenges encountered when working with web archives for these participants (n=25) include:

- Inconsistencies and incompleteness (r=11)
- Legalities for acquisition/access (r=8)
- Technical challenges (r=8)
- Challenges with learning new skills (r=6)
- Financial challenges (r=4)
- Producing documentation/metadata (r=2)
- Volume of data (r=2)
- Institutional challenges (r=1)
- Conceptual challenges (r=1)
- In-vivo representations (r=1)

In terms of workarounds and solutions for overcoming challenges, 5 participants provided free text responses, which were coded in four thematic representations including, challenges with learning new skills (r=4), volume of data (r=1), and broken links to files (r=1). These representations are further detailed below.

|> Challenges with learning new skills (r=4)

(r1)

- challenge: “learning curve was steep.”
- solution: “still working around that. asking a lot of questions of colleagues, attend conferences, reading documentation.”

(r2)

- challenge: “Learning how to use research tools (from a non-technical user's perspective).”

- solution: “attend lots of great workshops and tutorials e.g. Archives Unleashed, GLAM Workbench/Jupyter notebooks, Looking at using new services e.g. LinkGate & Solrwayback. Joining working groups with researchers (WARCnet e.g.) has been invaluable for learning from practitioners who are already actively using web archives for their research”

(r3)

- challenge: “Need to learn a lot about what web archives are and the technology that is used to create, curate and maintain them.”
- solution: “To overcome, working with colleagues in my institution, 'learning by doing', IIPC engagement, staff training”

(r4)

- challenge: “Limited technical skills to analyse the WARC-files and the information within them.”
- solution: “Attending one of the Archives Unleashed Toolkit's datathons was of help, but the downside was that it works best with WARC files created with Archive-It to which our library doesn't have a subscription.”

|> Broken links to files (r=1)

(r1)

- challenge: “Some problems are the fact that PDFs link to in a webpage are not accessible”
- solution: “the workaround involved trying variations of the URLs to see if I can stumble into the PDF somewhere. I would say the success rate is 25%, at best. But that is better than nothing”

|> Volume of data (r=1)

(r1)

- challenge: “The size of the collections and the difficulty of narrowing down a set of data that is manageable and appropriate”
- solution: “focus on smaller, curated collections”

Table 4.14: Thematic representation of responses for challenges encountered when working with web archives, by participants who identified with Library, Archive, or Web Archive environment (n=25)

| Theme representation for challenges encountered when working with web archives, by participants who identified with Library, Archive, or Web Archive environment (n=25) | No. of representations (R=44) |
|---|---|
| \|> Inconsistencies and incompleteness<br>● Broken links to files (e.g. PDFs, Excel etc.) (r=3)<br>● Erroneous crawls (r=3)<br>○ r: "Incomplete or erroneous crawls"<br>○ r: "The harvest is not always totally fine"<br>○ r: "when it gives errors in the capture"<br>● Layout/visual deficiencies (r=2)<br>○ r: "Sometimes the images are blurred"<br>○ r: "the visualization is not always right"<br>● Capturing dynamic content (r=1)<br>○ r: "the shallow delivery of dynamic content due to the limitations of the bots."<br>● Inconsistency with crawl frequency of early websites (r=1)<br>● r: "Variation in what is collected over time" (r=1) | r=12 |
| \|> Technical challenges<br>● Challenges to save sites due to firewall/security (r=1)<br>● Data storage (r=1)<br>● Data processing (r=1)<br>○ r: "Since I am interested in knowing about the entire archive, it means I am interested in multiple Petabytes of data, several million WARC files and Terabytes of index files. The largest barrier has been [the] ability to process this data."<br>● Difficult to create bulk data sets/share with researchers (r=1)<br>● File format obsolescence (r=1)<br>● Lack of IT infrastructure (r=1)<br>● Search and discovery challenges (r=1)<br>● Technical challenges (in general) (r=1) | r=8 |
| \|> Legalities for acquisition/providing access<br>● Challenges to provide access due to legislation, copyright and GDPR (r=5)<br>● Acquisition challenges for selective archiving (r=2)<br>○ Challenges to get permissions (r=1)<br>○ Acquisition restrictions for selective archiving (r=1) | r=8 |

| | |
|---|---|
| ● Embargoes (r=1) | |
| \|> Challenges with learning new skills<br>● r: "complexity of the WARC files"<br>● r: "It was a bit strange at first because I didn't have much of an idea of web archiving since I was more used to working with paper. But in a short time I got up to speed"<br>● r: "Learning how to use research tools (from a non-technical user's perspective)"<br>● r: "Limited technical skills to analyse the WARC-files and the information within them"<br>● r: "learning curve was steep"<br>● r: "Need to learn a lot about what web archives are and the technology that is used to create, curate and maintain them" | r=6 |
| \|> Financial challenges<br>● Cost of storage (r=1)<br>● Cost of services (r=1)<br>● Attaining funding (r=1)<br>● r: "On-premises access to web archives makes them economically inaccessible." | r=4 |
| \|> Documentation/metadata<br>● r: “confusing records”<br>● r: “Trying to guess the date when the site may have been crawled and when changes happen” | r=2 |
| \|> Volume of data<br>● r: “The size of the collections and the difficulty of narrowing down a set of data that is manageable and appropriate”<br>● r: “scale of the archive” | r=2 |
| \|> Conceptual challenges | r=1 |
| \|> Institutional challenges<br>● r: “a barrier can be institutional in convincing other areas of the organization about the value of the web archive and allocating funds to this type of work.” | r=1 |
| \|> In-vivo representations<br>● r: “Having access to the raw data, as a web archivist, is very beneficial“ | r=1 |

#### 4.3.5.2 Scholar, academic, lecturer, student, or IT/web design environment

In relation to challenges and participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment, 12 participants provided free text responses. 3 participants indicated that they encountered no/minimal challenges to using web archives for their research. 3 representations are in-vivo and offer other interpretations.

Table 4.15 offers an overview and breakdown of representations for challenges encountered when working with web archives for these participants (n=9) which includes:

- Inconsistencies and incompleteness (r=10)
- Legalities on access, use, and storage (r=8)
- Challenges with learning new skills (=7)
- Research methods and approaches (r=5)
- challenges in an IT/Business/Administrative environment (r=2)
- Lack of documentation/metadata (r=2)
- Volume of data for research (r=2)
- Performance related issues (r=1)
- In-vivo representations (r=3)

In terms of workarounds and solutions for overcoming challenges, 2 participants provided free text responses as outlined below.

|> Lack of documentation (r=1)

(r1)

- challenge/solution: “Trying to overcome issues relating to the lack of documentation by establishing close collaborations with curators and IT specialists at the archive”

|> Access, volume of data, inability to download data, lack of archival context (r=1)

(r1)

- challenge: “Closed access, volume, inability to download data, lack of archival context”
- solution: “still working on overcoming these, but working with specialist archival staff was essential.”

Table 4.15: Thematic representation of responses for challenges encountered when working with web archives, by participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=9)

| Theme representation for challenges encountered when working with web archives, by participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=9) | No. of representations (R=39) |
|---|---|
| \|> Inconsistencies and incompleteness<br>● Inconsistent in terms of what was saved (r=6)<br>○ r: “in terms of content: sometimes the website or the entry I am looking for is not archived”<br>○ r: “Many websites are hardly accessible, not enough material saved.”<br>○ r: "Missing image files”<br>○ r: "Broken links”<br>○ r: "inconsistent in terms of what was saved”<br>○ r: "inaccessible website”<br>● Inconsistent temporal coverage (r=2)<br>○ r: “Incomplete temporal coverage“<br>○ r: “inconsistent in terms of what was saved and when”<br>● Layout/visual deficiencies (r=1)<br>○ r: “Incorrect layout (in relation to live web)”<br>○ r: “Incompleteness in the data itself” | r=10 |
| \|> Legalities on access, use, and storage<br>● Legal challenges regarding access to data (r=4)<br>● Legal challenges regarding use of data (r=2)<br>● Inability to download data (r=1)<br>● Legal challenges regarding storage of data (r=1) | r=8 |
| \|> Challenges with learning new skills<br>● Having to acquire new programming skills (=2)<br>● Challenges with tools for web archive research (=1)<br>● Learning about the limitations of replay interfaces (=1)<br>● Difficulties to understand how web archives are set up (=1)<br>● Learning what a WARC file was (=1) | r=7 |
| \|> Research methods and approaches<br>● Lack of research methods/theory (r=2)<br>● r: “It is extremely difficult to put websites in the broader context of how they were used. And especially, because digital [quantitative] methods are prevailing over qualitative in the field Web History” | r=5 |

| | |
|---|---|
| ● r: “[research] community doesn't have enough [epistemological] assessment of web archives as historical sources yet. And this is crucial for interpretation.”<br>● Archived web as a source for research (r=1)<br>● r: “Gaining a proper understanding of archived web as a specific type of source and the consequences of these characteristics for [research] using archived web”<br>● Combining traditional methods with web archive research (r=1)<br>● r: “We had to think about ways to triangulate our insights, which is not always possible - we were working with interviews, html code and analogue media to do this.”<br>● Data analysis (r=1)<br>● r: “limited analytic functionality in web- based access interfaces” | |
| \|> Challenges in an IT/Business/Administrative environment<br>● r: “Funding and low awareness from stakeholders”<br>● r: “Dependency on a not-for-profit, third-party archiving initiative to meet our business needs [...] my company has not yet recognized the need for our own web archiving practice.“ | r=2 |
| \|> Lack of documentation/metadata<br>● r: “issues relating to the lack of documentation”<br>● r: “lack of archival context” | r=2 |
| \|> Volume of data for research<br>● r: “volume “<br>● r: “Working with large-scale data” | r=2 |
| \|> Performance related issues | r=1 |
| \|> In-vivo representations<br>● r: “One of the big barriers was getting started”<br>● r: “once I wanted to get more involved, who to contact!”<br>● r: “Too many to count!” | r=3 |

### 4.3.6 Skills and knowledge, before starting with web archives

Participants were asked about the useful skills or knowledge they had ‘Before’ they started their research in web archives. They were provided with a Likert scale, several answer options, and asked to tick all that applies. The Likert scale was organised as 3 levels of knowledge in terms of ‘a LOT of knowledge’, ‘SOME knowledge’ or ‘NO knowledge’.

Table 4.16 provides a representation of participant responses for this section. All participants (N=44) responded to this section, and some observations are outlined below.

In terms of having 'a LOT of knowledge' some participants identified with the following:

- Excel (or other spreadsheet) - Intermediate/Advanced (n=19)
- How websites are built/ made/ updated (n=16)
- How Fair Use works - copyright, reproduction rights, fair use (n=14)
- How digital legal deposit works and what it is (n=14)
- How digital curation works - collection, metadata, storage, access, long-term preservation (n=12)

In terms of having 'SOME knowledge' some participants identified with the following:

- How the internet works - Geo-IP, servers, browsers, domains, hosting etc. (n=30)
- How digital curation works - collection, metadata, storage, access, long-term preservation (n=24)
- Excel (or other spreadsheet) - Intermediate/Advanced (n=21)
- How Fair Use works - copyright, reproduction rights, fair use (n=21)
- Database creation and maintenance (n=20)
- How websites are built/ made/ updated (n=20)
- Metadata analysis (n=20)

In terms of having NO knowledge' some participants identified with the following:

- Python - Basic/intermediate (n=32)
- Java - Basic/intermediate (n=38)
- HTTrack (n=37)
- How web archiving works - WARCs, Capture tools, storage, and playback (n=20)
- Data analysis, such as topic modelling, textual analysis, etc. (n=18)
- How digital legal deposit works and what it is (n=17)

Table 4.16: Representation of participant responses for the skills and knowledge they had 'Before' they started their research with web archives (N=44)

| Answer Choices for skills and knowledge which proved to be useful | Yes - I had a LOT of knowledge | Yes - I had SOME knowledge | No - I had NO knowledge |
|---|---|---|---|
| How websites are built/made/updated (N=44) | n=16 | n=20 | n=8 |
| How the internet works - Geo-IP, servers, browsers, domains, hosting etc. (N=44) | n=8 | n=30 | n=6 |
| How web archiving works - WARCs, Capture tools, storage, and playback (N=44) | n=9 | n=15 | n=20 |
| How digital curation works - collection, metadata, storage, access, long-term preservation (N=44) | n=12 | n=24 | n=8 |
| How Fair Use works - copyright, reproduction rights, fair use (N=44) | n=14 | n=21 | n=9 |
| How digital legal deposit works and what it is (N=44) | n=14 | n=13 | n=17 |
| Excel (or other spreadsheet) - Intermediate/Advanced (N=44) | n=19 | n=21 | n=4 |
| Data analysis, such as topic modelling, textual analysis, etc. (N=44) | n=7 | n=19 | n=18 |
| Metadata analysis (N=44) | n=10 | n=20 | n=14 |
| Database creation and maintenance (N=44) | n=9 | n=20 | n=15 |
| Python - Basic/intermediate (N=44) | n=1 | n=11 | n=32 |
| Java - Basic/intermediate (N=44) | n=2 | n=4 | n=38 |
| HTTrack (N=44) | n=1 | n=6 | n=37 |

### 4.3.7 Other useful skills and knowledge, before starting with web archives

Provided with a comment box, participants were further asked to describe any 'Other' skills or knowledge they had before they commenced working/researching with web archives. 20 participants provided free text responses, which were coded into several thematic

representations. The responses for this section are analysed through the number of times a particular skill or knowledge is mentioned and is documented as a representation (R/r=).

Table 4.17 offers an overview and breakdown of such representations which include:

- Research methods/approaches (r=9)
- Information sciences (information studies) (r=7)
- Programming, scripting languages (r=6)
- Data analysis skills (r=4)
- Website design/browser developer tools (r=4)
- Finding information/services (r=3)
- Software and tools (r=3)
- Languages/translation skills (r=2)
- No skills (r=2)
- Graphic design skills (r=1)
- Social media skills (r=1)
- Skills in usability studies (r=1)

Table 4.17: Thematic representation of participant responses for 'Other' skills they had before starting their research with web archives which proved useful (n=20)

| Theme representation for 'Other' useful skills they had before starting their research with web archives (n=20) | No. of representations (R=43) |
|---|---|
| \|> Research methods and approaches<br>● Analytical thinking (r=2)<br>● Historical research skills/methods (r=2)<br>● Archival research skills (r=1)<br>● Digital humanities skills/methods (r=1)<br>● Mathematics (r=1)<br>● Understanding of provenance (r=1) | r=8 |
| \|> Information sciences (information studies)<br>● Archiving PDF/Screenshot, type of web archiving (r=1)<br>● Data management skills (r=1)<br>● Document database management systems (r=1)<br>● Library information science (r=1)<br>● Media formats (r=1)<br>● Preserving net art (r=1)<br>● Records management (r=1) | r=8 |

| | |
|---|---|
| ● Semantic web technologies for digital libraries (r=1) | |
| \|> Programming, scripting languages<br>● Programming tools (in general) (r=2)<br>● JavaScript (r=1)<br>● Perl (r=1)<br>● PHP (r=1)<br>● Unix shell (r=1) | r=6 |
| \|> Data analysis skills<br>● Visual / multimodal analysis skills (r=2)<br>● Pre-processing data (r=1)<br>● Semiotic analysis skills (r=1) | r=4 |
| \|> Website design/browser developer tools<br>● Browser developer tools (r=1)<br>○ r: “optimizing use of browsers' dev tools”<br>● Website design (r=3)<br>○ Web design (in general) (r=1)<br>○ r: “Looking at websites as objects (some static, some changing) helped in grasping web archiving conceptually.”<br>○ r: “a background creating flash and CSS websites“ | r=4 |
| \|> Finding information/services<br>● r: “trying different keywords, URLs, thinking about the way information in an organization might be organized.”<br>● r: “some training in finding things in libraries” | r=3 |
| \|> Software and tools<br>● Maths tools (r=1)<br>● MySQL (r=1)<br>● Statistical tools (r=1) | r=3 |
| \|> Languages/translation skills | r=2 |
| \|> No skills | r=2 |
| \|> Graphic design skills | r=1 |
| \|> Social media skills | r=1 |
| \|> Skills in usability studies | r=1 |

## 4.3.8. Other useful skills or knowledge participants 'WISH' they had

Provided with a comment box, participants were asked about other useful skills that they 'WISH' they had before they started their research in web archives. 18 participants provided free text which was coded into several thematic representations. 5 representations are in-vivo and offer other interpretations. The responses for this section are analysed through the number of times a particular skill or knowledge is mentioned and is documented as a representation (R/r=).

Table 4.18 offers an overview, and breakdown of such thematic responses which include:

- Software and tools (r=7)
- Web design/internet related skills (r=7)
- Programming, scripting languages (r=5)
- Finding information/services (r=2)
- Application of metadata (r=1)
- Collaborative skills (r=1)
- Digital legal deposit (r=1)
- Ethnography (r=1)
- Glossary of terminology (r=1)
- Managing protected data (r=1)
- Marketing and public relations (r=1)
- In-vivo representations (r=5)

Table 4.18: Thematic representation of participant responses for other useful skills or knowledge they 'WISH' they had before they started their research in web archives (n=18)

| Theme representation for other useful skills or knowledge they 'WISH' they had before they started their research in web archives (n=18) | No. of representations (R=33) |
|---|---|
| \|> Software and tools<br>● Data extraction, cleaning, and management (r=3)<br>○ Data cleaning tools (r=1)<br>○ Excel (or other spreadsheet) (r=1)<br>○ Regular expressions/Regex (r=1)<br>● Distributed processing (r=2)<br>○ Hadoop (r=1)<br>○ Spark (r=1)<br>● Computing infrastructure (r=1)<br>○ Amazon Web Services (r=1) | r=7 |

| | |
|---|---|
| ● Crawling software (r=1)<br>○ Heritrix: basic-advanced profile knowledge for functionalities (r=1) | |
| \|> Web/internet related skills<br>● Web design/development (r=4)<br>○ Web design/development tools (=1)<br>○ Understanding of HTML (r=1)<br>○ r: “Understanding how websites have been built over the past 30+ years.”<br>○ r: “How websites are built/ made/ updated”<br>● Better understanding of the technical history of the web (r=1)<br>● Better understanding of technical history of the internet (r=1)<br>● How the internet works (r=1) | r=7 |
| \|> Programming, scripting languages<br>● Programming (r=2)<br>○ Programming (in general) (r=1)<br>○ r: “if only I had some previous programming knowledge before starting my research. It would have been really useful throughout my research and archiving job.”<br>● R (r=2)<br>● Python (r=1) | r=5 |
| \|> Finding information/services<br>● r: “A list of more web archives”<br>● r: “topical knowledge about where to look” | r=2 |
| \|> Application of metadata<br>● r: “Information on how best to assign metadata” | r=1 |
| \|> Collaborative skills<br>● r: “How to collaborate with others” | r=1 |
| \|> Digital legal deposit<br>● r: “How digital legal deposit works and what it is” | r=1 |
| \|> Ethnography | r=1 |
| \|> Glossary of terminology<br>● r: “A glossary of terminology would also be helpful” | r=1 |
| \|> Managing protected data<br>● r: “Handling protected data (sensitive data and copyright protected data)” | r=1 |

| | |
|---|---|
| \|> Marketing and public relations | r=1 |
| \|> In-vivo representations<br>● r: “how indexes are generated, what they contain, and the potential uses they can be put to“<br>● r: “(hyper)link tracing / retrieval would be useful”<br>● r: “I really use web archives in a limited capacity and I am not trying to get too fancy.”<br>● r: “All the necessary skills were provided by the [web archive] team”<br>● r: “Sustainability (long-term availability) of the Internet Archive’s Wayback Machine” | r=5 |

### 4.3.9 New skills acquired through curation/use of web archives

Provided with a comment box, participants were asked to provide some examples of new skills they learned AFTER starting their research in web archives. 22 participants provided free text which was coded into several thematic representations. 2 representations are in-vivo and offer other interpretations. The responses for this section are analysed through the number of times particular skills are mentioned and are documented as a representation (R/r=).

Table 4.19 offers an overview, and breakdown of such thematic responses which include:

- Web archives, web archiving, curation (r=21)
- Software and tools (r=18)
- Digital curation processes/workflows (r=17)
- Data analysis skills (r=9)
- Programming/scripting languages (r=7)
- Web/internet related skills (r=3)
- Research methods and approaches (r=3)
- Database creation and maintenance (r=1)
- Digital legal deposit (r=1)
- Fair use, copyright, reproduction rights (r=1)
- Managing protected data (r=1)
- In-vivo representations (r=2)

Table 4.19: Thematic representation of participant responses for new skills or knowledge acquired after starting their research in web archives (n=19)

| Theme representation of responses for new skills or knowledge acquired after starting research in web archives (n=22) | No. of representations (R=84) |
|---|---|
| \|> Web archives, web archiving, curation<br>● How web archiving works (r=17)<br>○ Understanding of web archiving tools (r=4)<br>○ Web archiving (in general) (r=3)<br>○ How crawling/capture works (r=2)<br>○ Understanding of data storage (r=2)<br>○ Understanding of playback/replay (r=2)<br>○ Understanding of WARCs (r=2)<br>○ How to create web archiving workflows (r=1)<br>○ How web archives are organised (r=1)<br>● Educational activities for web archiving (r=1)<br>● International collaboration on web archiving (r=1)<br>● Web archiving standards (r=1)<br>● Other representation (r=1)<br>○ r: "Implementing foreign professional concepts into our own web archiving practice." | r=21 |
| \|> Software and tools<br>● Data extraction, cleaning, and management (r=5)<br>○ Excel, spreadsheets (r=3)<br>○ Regex/ Regular expressions (r=1)<br>○ r: "Tools for data cleaning"<br>● Crawling software (r=2)<br>○ Heritrix (r=2)<br>● Network analysis (r=3)<br>○ Gephi (r=3)<br>● Curating collections: selection, configuring and scheduling crawls, annotating seeds, performing QA (r=2)<br>○ CWeb (r=1)<br>○ NetArchiveSuite (r=1)<br>● Distributed processing (r=2)<br>○ Hadoop (r=1)<br>○ Spark (r=1)<br>● Replaying archived web data (r=1)<br>○ Open Wayback (r=1)<br>● Web archive access and analysis | r=18 |

| | |
|---|---|
| ○ GLAM Workbench (Jupyter Notebooks) (r=1)<br>● Computing infrastructure (r=1)<br>○ Amazon Web Services (AWS) (r=1)<br>● r: "using dev tools" | |
| \|> Digital curation processes/workflows<br>● Metadata (r=6)<br>● Long-term preservation/infrastructures (r=3)<br>● Access (r=2)<br>● Collection (r=2)<br>● Digital storage (r=2)<br>● How digital curation works (r=2) | r=17 |
| \|> Data analysis skills<br>● Data analysis (in general) (r=3)<br>● Link analysis (r=1)<br>● Quantitative data analysis (r=1)<br>● Qualitative data analysis (r=1)<br>● Text analysis (r=1)<br>● Visual analysis (r=1)<br>● Large-scale data analysis (r=1)<br>○ r: “Understanding better the challenges and potential for large-scale data analysis.” | r=9 |
| \|> Programming/scripting languages<br>● Programming and visualisations with R (r=4)<br>● Python scripts/libraries (r=2)<br>● Shell scripting (r=1) | r=7 |
| \|> Web/internet related skills<br>● r: “Above all, how the creation of the Web works and behaves in general”<br>● r: “How websites are updated”<br>● r: “How the internet works - Geo-IP, servers, browsers, domains, hosting etc. “ | r=3 |
| \|> Research methods and approaches<br>● r: “Knowing more about research uses of archived web”<br>● r: “theoretical approaches to web archives and source code.”<br>● r: “how to keep notes about where information/data comes from” | r=3 |
| \|> Database creation and maintenance | r=1 |

| | |
|---|---|
| \|> Digital legal deposit<br>● r: "How digital legal deposit works and what it is" | r=1 |
| \|> Fair use, copyright, reproduction rights<br>● r: "How Fair Use works - copyright, reproduction rights, fair use" | r=1 |
| \|> Managing protected data<br>● r: "Handling protected data" | r=1 |
| \|> In-vivo responses<br>● r: "It is hard to list as I would say that I have a fairly advanced knowledge of the computational aspects of working with WARCs at scale, and knew almost nothing starting out."<br>● r: "Most of my digital skills!" | r=2 |

### 4.3.10 Changes in research questions or parameters

Provided with three multiple choice options, participants were asked if their research question or parameters changed after starting their research project(s), including the disruptions caused by the COVID pandemic.

Figure 4.4 provides an overview of participant responses (N=44) and indicates the following:

- No – they did not change (43.18%, n=19)
- Yes – they changed a little (29.54%, n=13)
- Yes – they changed a lot (27.27%, n=12)

Further to this, participants who answered 'Yes' were asked to describe how their research question or parameters changed in a comment box. 19 participants provided free text responses which were coded into several thematic representations. 5 representations are in-vivo and offer other interpretations. The responses for this section are analysed through the number of times changes to research questions or parameters are mentioned and are documented as a representation (R/r=).

Table 4.20 provides of an overview of such representations which include changes in research questions or parameters that are related to:

- Research methods/approaches (r=11)
- Web archives, web archiving, curation (r=8)
- In-vivo representations (r=5)

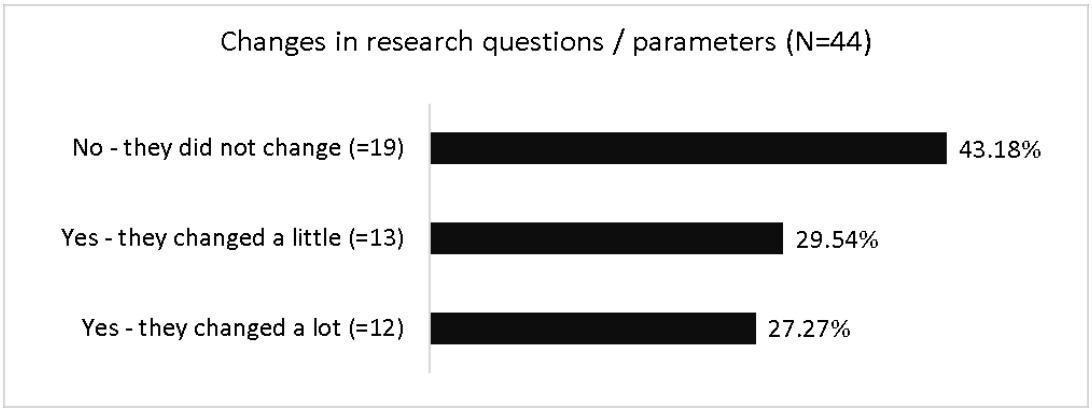


Figure 4.4: Representation of participant responses for changes in research questions or parameters (N=44)

Table 4.20: Thematic representation of participant responses for changes to research questions or parameters (n=19)

| Theme representation of responses for changes to research questions or parameters (n=19) | No. of representations (R=24) |
|---|---|
| \|> Research methods/approaches<br>● Data analysis, data cleaning (r=4)<br>○ r: "a recurring theme when working with large amounts of archived web data is discovering new issues with the data which require redoing analyses, often with additional data cleaning involved"<br>○ r: "The basic research question and purpose remained the same (learning about the archive in order to give better care to the items), but choosing to analyze the derivative crawl data and the CDX index files changed the types of questions asked of the data. I went in thinking it would be a lot more detailed, but found it better to start at higher levels with derivative data and metadata before going in deeper with data held in the WARCs."<br>○ r: "The opportunities and tools available for large-scale data analysis has changed quickly during the time I have worked with web archives"<br>○ r: "I always find that digging into some data gives me new ideas for new things I can dig out."<br>● Access to raw data (r=1) | r=11 |

| | |
|---|---|
| ○ r: “I initially thought it might be possible to get the raw data - WARC files - from the various libraries but that was not the case, so derived data or seedlists were used instead”<br>● Attendance of online conferences/webinars (r=1)<br>○ r: “I could not participate [in] on-site conferences, however I could participate online on various webinars, conferences I could not afford to participate on-site. These events have broadened my research perspectives and I could add some more analyzing aspects to my phd project.”<br>● Blog design/communities (r=1)<br>○ r: “I realised that changes in the design of blogs that were not visible in the integrated blog archive were usually maintained in the archived versions of the blog and that the 'same' web object changed over time. This allowed me to make connections with the bloggers' identity transformation and belonging over time, which in turn meant I changed my methodology from a purely contemporary analysis to one which involved recent history.”<br>● COVID disruption (r=1)<br>○ r: “Covid has disrupted travelling to individual libraries to consult datasets.”<br>● Data centred approach (r=1)<br>○ r: “Completely new data centered approach”<br>● Digital humanities tools/methods/approaches (r=1)<br>○ r: “Digital Humanities and using large scale computation methods and tools like Hadoop/Spark through R with Jupyter Notebooks and other similar tools“<br>● Statistical analysis requirements (r=1)<br>○ r: “The requirement for better knowledge of using spreadsheets in [statistical] analysis “ | |
| \|> Web archives, web archiving, curation<br>● Collection development strategies/decisions (r=4)<br>○ r: "collaborative archiving”<br>○ r: "My interest is in how collections can be created and communicated. This has changed a lot, with much more emphasis on working collaboratively to build collections.”<br>○ r: “I didn't know anything about web archiving until I tried Conifer myself. I've watched demos for ArchiveIt. Now that I've done the archiving I understand the practice of using some of these tools, which helps in making decisions for future collecting decisions.” | r=8 |

| | |
|---|---|
| ○ r: “At the beginning, more administrative-type pages were collected, later it was expanded to more cultural topics.“<br>● Challenges with social media archiving (r=2)<br>● Learning automation processes (r=1)<br>● Priorities change (r=1)<br>○ r: “times have expanded and interest was no longer a priority” | |
| \|> In-vivo representations<br>● r: “I find that with every project, the more you learn, the more you refine the question and the parameters for the search.”<br>● r: “I'm a reference librarian, so my research projects are always changing.”<br>● r: “It has been a process of constant development, since I have not been bound into a clearly bounded project as such.”<br>● r: “looking at specific types of written sources”<br>● r: “Often I am working with a client, so when we learn that certain information is not available, we can refine the question and be more targeted in what we do look for” | r=5 |

## 4.4 Citation Practises

In this section we look at referencing styles and practises, and the challenges for the citation of archived web content and datasets of archived web content.

### 4.4.1 Referencing styles and practises

Participants were asked about the referencing systems they use for citing sources in their research in general, when using materials other than web archives. They were offered a list of choices and asked to tick all that applied. They were also offered the option of ‘Other’ to enter free text.

Figure 4.5 offers a representation of participant responses (N=44) and indicates the following:

- APA (American Psychological Association) (34.09%, n=15)
- MLA (Modern Languages Association) (27.27%, n=12)
- Harvard system (18.18%, n=8)
- IEEE (Institute of Electrical and Electronics Engineers) (6.81%, n=3)
- MHRA (Modern Humanities Research Association) (2.27%, n=1)
- Other (50%, n=22)

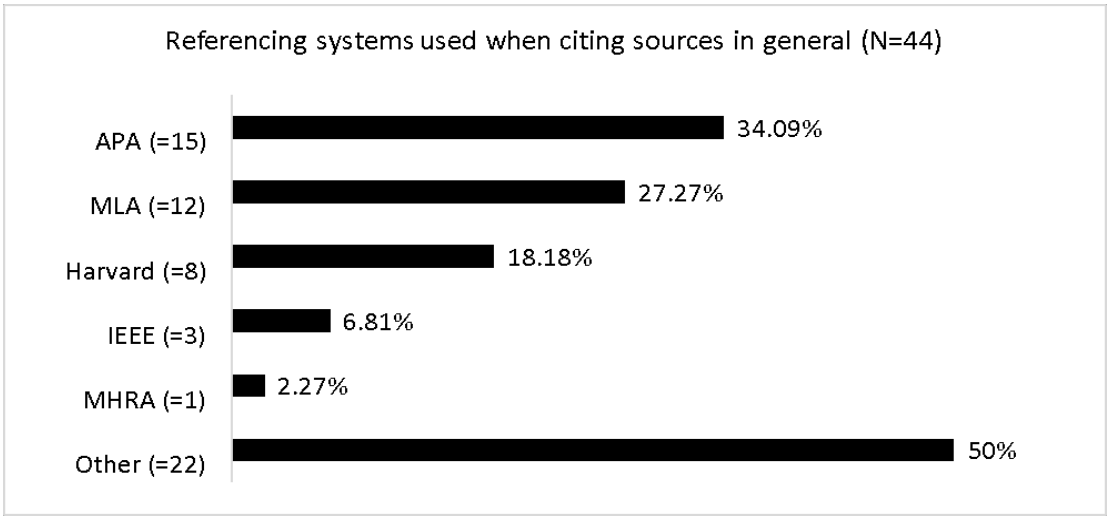


Figure 4.5: Representation of participant responses for referencing systems used when citing sources in general (N=44)

In addition, 22 participants entered free text responses for 'Other' referencing systems. The responses were coded into several theme representations. 2 representations are in-vivo and offer other interpretations. The responses for this section are analysed through the number of times referencing systems or standards are mentioned and is documented as a representation (R/r=).

Table 4.21 offers an overview, and breakdown of the thematic representations which include:

- Other referencing styles
- Other standards/specifications
- Non-applicable for some participants (r=4)
- Depends on journal/publisher/proceedings (r=2)
- Internal/institutional citation formats (r=2)
- Reference management applications/mark-up (for any style) (r=2)
- In-vivo representations (r=2)

Table 4.21: Thematic representation of participant responses for 'Other' referencing systems used (n=22)

| Theme representation for 'Other' referencing systems used (n=22) | No. of representations (R=25) |
|---|---|
| \|> Other referencing styles<br>● Chicago (r=6)<br>● Turabian (r=1) | r=7 |
| \|> Other standards/specifications<br>● ISO standards (r=2)<br>● Digital Object Identifier (r=1)<br>● r: "Use DOIs to cite datasets where they exist. (e.g. UK Web Archive derived datasets)"<br>● ISBD (International Standard Bibliographic Description)(r=1)<br>● RDA (Resource Description and Access) (r=1)<br>● ГОСТ (GOST) (r=1) | r=6 |
| \|> Non-applicable for some participants | r=4 |
| \|> Depends on journal/publisher/proceedings | r=2 |
| \|> Internal/institutional citation formats<br>● r: "Tend to use an internal format"<br>● r: "Our institutional citation formats are unique and varied" | r=2 |
| \|> Reference management applications/mark-up (for any style)<br>● Zotero (r=1)<br>● LaTeX/BibTex (r=1) | r=2 |
| \|> In-vivo representation<br>● r: "I haven't written academic papers citing web archives (generally, I write policy papers that are about web archiving)"<br>● r: "Not yet published" | r=2 |

## 4.4.2 Challenges for citing archived web content

Participants were asked if they have any challenges when citing archived web content from a web archive. They were provided with three answer choices of 'Yes', 'No', or 'Sometimes'.

Figure 4.6 provides an overview of participant responses (N=44) which indicates the following:

- No (52.27%, n=23)
- Sometimes (36.36%, n=16)
- Yes (11.36%, n=5)

Table 4.22 offers a breakdown of the results in line with the participants position and indicates that there is no relevant pattern or differentiation between one community of practice or the other.

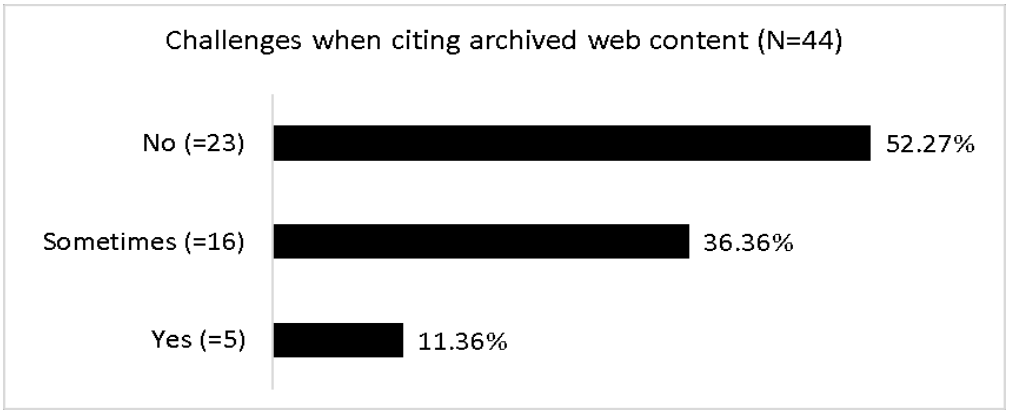


Figure 4.6: Representation of participant responses for challenges when citing archived web content (N=44)

Table 4.22: Representation of participant responses (by position) for challenges when citing archived web content from a web archive (N=44)

| \|> Library, Archive, or Web Archive environment (n=30) | \|> Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=14) |
|---|---|
| • No (n=15)<br>• Sometimes (n=11)<br>• Yes (n=4) | • No (n=8)<br>• Sometimes (n=5)<br>• Yes (n=1) |

Participants who selected ‘Yes’ or ‘Sometimes’ were further asked to describe some of the challenges they have for citing archived web content in a comment box. 20 participants provided free text responses which were coded into several thematic representations. 4 representations are in-vivo and offer other interpretations.

Table 4.23 offers an overview and breakdown of such representations which includes:

- Lack of guidelines/standards/best practice (r=7)
- Challenges with citing content from legal deposit/archives with restrictive access (r=4)
- Uncertainties for citing archived web content (r=4)
- Challenges specific to the URL for archived web content (r=2)
- Not easy to cite sources from a web archive (in general) (r=2)
- Problem to find dates/creators for the websites in a web archive (r=1)
- In-vivo representations (r=4)

Table 4.23: Thematic representations of participants’ descriptions for challenges when citing archived web content (n=20)

| Theme representations for challenges when citing archived web content (n=20) | No. of representations (R=24) |
|---|---|
| \|> Lack of guidelines/standards/best practice<br>● Lack of guidelines (in general) (r= 1)<br>● r: “Agreeing on best practice”<br>● r: “Lack of rules for citing 'popular' things like forums (or more recently, but less likely to be archives, social media)”<br>● r: “Sometimes it is not quite clear what the best way to cite a source is.”<br>● r: “Referencing standards are sometimes not adapted to the archival materials.”<br>● r: “The existing systems don't have a model for this type of content.”<br>● r: “referencing system doesn't give a clear guideline for digital sources in general“ | r=7 |
| \|> Challenges with citing content from legal deposit/archives with restrictive access<br>● r: “Citing historic content in a closed archive only accessible by other researchers in a [persistent] way<br>● r: “Copying and pasting a URL from a reading room viewer is not possible as the browsers are locked down.” | r=4 |

| | |
|---|---|
| ● r: “I am aware that there are challenges for users of web archives. Some of these are a result of regulatory restrictions (eg it's not easy to copy and paste urls).”<br>● r: “The basic problem is, that if you want to cite to some elements that are in a collection with restricted access, nobody beyond your institution affiliation can check your links. Furthermore in some case a special knowledge required either way to retrieve data from WARC files.” | |
| \|> Uncertainties for citing archived web content<br>● Should it be cited like a normal website? (r=1)<br>○ r: “It is difficult to know if you should cite it similar to a website”<br>● Should the source be treated as a normal URL? (r=1)<br>○ r: “Unsure whether to treat it is a URL”<br>● Should the web archive be acknowledged? (r=1)<br>○ r: “whether the archive should be acknowledged”<br>● What dates should be used? (r=1)<br>○ r: “what dates should be used (capture date, access date, date of original publication, e.g. a blog post or article).” | r=4 |
| \|> Challenges specific to the URL for archived web content<br>● r: “The standard URL identifier derived from Wayback, while adequate, is unwieldy and not easily read by humans.“<br>● r: “Ensuring stability of references, even if archive systems change“ | r=2 |
| \|> Not easy to cite sources from a web archive (in general)<br>● r: “Web Archives tend not to offer an easy way to generate a citation.”<br>● r: “It is not easy to cite parts of website from web archive” | r=2 |
| \|> Problem to find dates/creators for the websites in a web archive<br>● r: “Finding dates for some archived sites, sure we can find technical metadata for when it was archived, but not always the original source creation, or even who precisely the creators and contributors may be.” | r=1 |
| \|> In-vivo representations<br>● r: “The web address is not stable”<br>● r: “Lengthy citations are of limited value to my colleagues in the private sector business I work in - they may not care about the details, but I want to provide thorough citations in case we need to go back to something.” | r=4 |

| | |
|---|---|
| • r: “References can be either incomplete, not cited correctly or incorrect which requires further research.”<br>• r: “We try to cite to institution-created sources. If we are not able to find an official source from our institution, we try to find a way to cite to an archived version that we think will be stable or to re-capture the information in an institutional product that will (hopefully) be stable over time.” | |

### 4.4.3 Challenges for citing datasets with archived web content

Participants were asked whether they have any challenges when citing datasets of archived web content. They were provided with the answer choices of ‘Yes’, ’No’, or Sometimes’, or could opt out from answering.

Figure 4.7 offers a representation of the participant responses (N=44), of which 8 participants (18.18%, n=8) provided no answer. The remaining 36 participants indicated the following:

- No (38.36%, n=17)
- Sometimes (27.27%, n=12)
- Yes (15.90%, n=7)

Further to this, participants who answered ‘Yes’ or ‘Sometimes’ were provided with a comment box and asked to describe some of the challenges they have with citing datasets of archived web content. 16 participants provided free text responses which were coded into several thematic representations.

Table 4.24 offers an overview and breakdown of such representations (n=16) which includes:

- Lack of guidelines/standards for citing datasets (r=5)
- Amount of data/details to include in a dataset citation (r=3)
- Not easy to cite datasets (in general) (r=3)
- Uncertainties for citing datasets with archived web content (r=2)
- Data/content reliability within a dataset (r=1)
- Incorporation of PWID in web archives as a citation aid (r=1)
- Preservation quality of datasets (r=1)
- System restrictions (r=1)
- In-vivo representations (r=2)

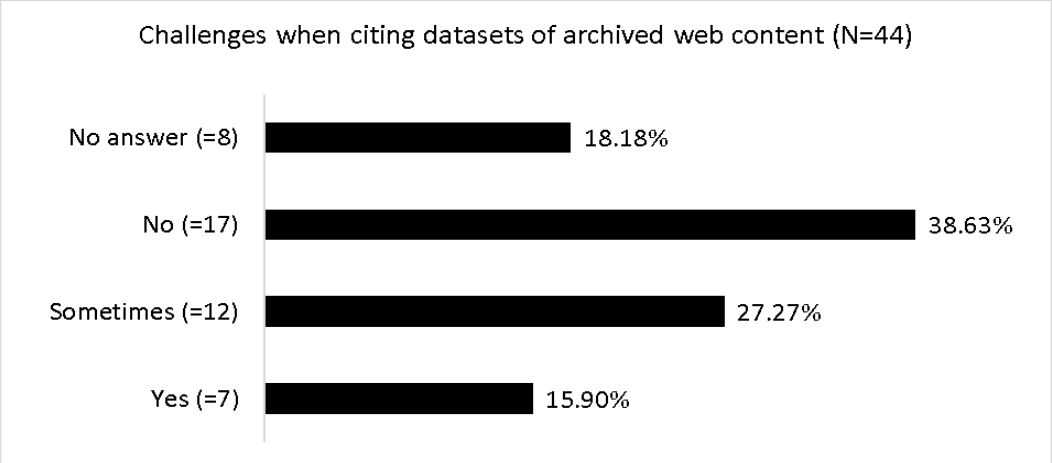


Figure 4.7: Representation of participant responses for citation challenges using datasets of archived web content (N=44)

Table 4.24: Thematic representation of participants' descriptions of challenges for citing datasets of archived web content (n=16)

| Theme representations for challenges when citing datasets of archived web content (n=16) | No. of representations (R=19) |
|---|---|
| \|> Lack of guidelines and standards for citing datasets<br>● r: "It's just hard to reference, there are almost no guidelines on the subject."<br>● r: "I don't know if there is an agreed standard for citing datasets."<br>● r: "Sometimes it is not quite clear what the best way to cite a source is."<br>● r: "The existing systems don't have a model for this type of content"<br>● r: "Referencing standards are sometimes not adapted to the archival materials" | r=5 |
| \|> Amount of data / details to include in a dataset citation<br>● r: "Amount of detail required is difficult to present in a manner that people can quickly scan and understand ."<br>● r: "Citing a large corpus that was extracted from [a web archive] with specific parameters, what do you preserve (the actual data, the methods/algorithms/filters/programs) ? - hard for others to redo the research without exact knowledge of the datasets."<br>● r: "How much to include in relation to describing how the data were collected - depending on context." | r=3 |

| | |
|---|---|
| \|> Not easy to cite datasets (in general)<br>● Not easy to cite datasets (r=2)<br>● r: “I think making references to datasets themselves is really problematic luckily I did not need it during my phd research.” | r=3 |
| \|> Uncertainties for citing datasets with archived web content<br>● Should the web archive be acknowledged? (r=1)<br>● What dates should be used? (r=1) | r=2 |
| \|> Data/content reliability within a dataset<br>● r: “There is also the issue of the 'page' and if information appears below the original landing page when scrolling down, for example” | r=1 |
| \|> Incorporating PWID in web archives as a citation aid | r=1 |
| \|> Preservation quality of datasets<br>● r: “Derived data sets from web archived data may not be properly preserved” | r=1 |
| \|> System restrictions<br>● r: “we don't always have ways of recording the source of web content in our systems.” | r=1 |
| \|> In-vivo representations<br>● r: “Unable to recall”<br>● r: “It is not a task that I do continuously” | r=2 |

## 4.5 Resources and Data Sharing

In this final section, we look at participants’ suggestions for useful resources. We further examine participants’ data sharing practices and the types of repositories they use for data sharing. The section ends with an outline of any final comments by participants.

### 4.5.1 Useful resources

Provided with a comment box, participants were asked to list any resources that they found useful to further their skills and knowledge in their research with web archives. For example, this could be an online or in-person training course, workshop, or mentorship. 30 participants provided free text responses which were coded into several thematic representations. 2 representations are in-vivo and offer other interpretations. The responses for this section are

analysed through the number of times an individual resource is mentioned and is documented as a representation (R/r=).

Table 4.25 offers an overview, and breakdown of the thematic representations which include:

- Training, workshops, courses (r=26)
- Software and tools (r=16)
- Websites, web pages, blogs (r=15)
- Collaborations and mentorship (r=14)
- Consortiums, networks, conferences (r=14)
- Introductions, guides, manuals (r=4)
- Literature (r=3)
- Information sciences (information studies) (r=1)
- Providing learner support (r=1)
- Self-learning (r=1)

Further to this, the same participants (n=30) mentioned several organisations, institutions, consortiums, projects, networks, and conferences (r=29) which they found useful as outlined below:

- International Internet Preservation Consortium (r=6)
- WARCnet (r=4)
- British Library, UK Web Archive (r=3)
- RESAW (r=3)
- Rhizome, Conifer, Webrecorder (r=3)
- Archives Unleashed (r=2)
- Digital Preservation Coalition (=1)
- German Literature Archive Marbach (r=1)
- Koninklijke Bibliotheek (r=1)
- National Digital Stewardship Residency for Art (r=1)
- Netarkivet/Aarhus University (r=1)
- Tara Repository (TCD), Ireland (r=1)
- The National Archives, UK (r=1)
- Trinity College Dublin, Ireland (r=1)

Table 4.25: Thematic representation of participant responses for useful resources to further their skills or knowledge in their research with web archives (n=30)

| Theme representations for useful resources to further skills and knowledge for research with web archives (n=30) | No. of representations (R=81) |
|---|---|
| \|> Training, workshops, courses<br>● International Internet Preservation Consortium (r=5)<br>○ r: "IIPC Congress and workshops about tools"<br>○ r: "IIPC webinars, workshops"<br>○ r: "Training course from the IIPC"<br>○ r: "IPC sponsored events"<br>○ r: "IIPC Webinar about Web Archive"<br>● Training from a web archive (r=3)<br>● Archives Unleashed Datathons (r=2)<br>● Institutional training/courses (r=2)<br>● Online training/tutoring (r=2)<br>● RESAW (r=2)<br>○ r: "workshop at RESAW conferences/meeting"<br>○ r: "There was a great web archiving hands-on tutorial that Jefferson Bailey and Vinay Goel ran at the Aarhus RESAW conference. It was incredibly useful."<br>● Training/courses (in general) (r=2)<br>● Workshops (in general) (r=2)<br>● MODE Summer School, UCL, Institute of Education, Knowledge (r=1)<br>○ r: "Multimodality Summer School (week-long at Institute of Education / Knowledge Lab)"<br>● Netlab, Aarhus University (r=1)<br>○ r: "Netlab - course by Aarhus University"<br>● Rhizome (r=1)<br>○ "[lecture] by Dragan Espenshied from Rhizome"<br>● The National Archives UK/ Digital Preservation Coalition (r=1)<br>○ r: "Novice to Knowhow from TNA and DPC"<br>● Training from a digital repository (r=1)<br>● Trinity College Dublin (r=1)<br>○ r: "Digital Humanities course run by Trinity College Dublin" | r=26 |
| \|> Software and tools<br>● Data analysis, cleaning, transformation (r=6)<br>○ Archives Unleashed Toolkit (r=2)<br>○ Excel (advanced) (r=1)<br>○ Pandas (r=1)<br>○ Power BI (r=1)<br>○ Tableau (r=1)<br>● Crawling software (r=3)<br>○ ArchiveWeb.page (r=1)<br>○ Conifer (prior, Webrecorder) (r=1) | r=16 |

| | |
|---|---|
| ○ Heritrix (r=1)<br>● Network analysis (r=3)<br>○ Gephi (r=2)<br>○ LinkGate (r=1)<br>● Information retrieval (r=2)<br>○ Solrwayback (r=2)<br>● Programming, scripting languages and computing environments (r=1)<br>○ Jupyter Notebooks (r=1)<br>● Web archive access and analysis (r=1)<br>○ GLAM Workbench (r=1) | |
| \|> Websites, web pages, blogs<br>● International Internet Preservation Consortium (r=7)<br>● Zenodo (r=2)<br>● ArchiveWeb.page (r=1)<br>● Conifer (r=1)<br>● Heritrix (r=1)<br>● One Terabyte of a Kilobyte Age (Blog) (r=1)<br>● Pandas (r=1)<br>● SolrWayback (r=1) | r=15 |
| \|> Collaborations and mentorship<br>● Library, Archive, or Web Archive environment (r=8)<br>○ Mentorship by library staff (r=3)<br>○ r: "brainstorming with team members"<br>○ r: "learning from colleagues”<br>○ r: "virtual meetings to discuss specific topics between all the personnel dedicated to the [web archive]"<br>○ r: "Working with colleagues who have a detailed knowledge of web archiving"<br>○ r: "Working with researchers using archived web data"<br>● Scholar, Academic, Lecturer, Post-grad/PhD, or working IT/Web Design environment (r=6)<br>○ r: "collaboration with web archives"<br>○ r: "Conversations with web archiving service providers and customers"<br>○ r: "learn a bit from [staff at archive]"<br>○ r: "Mentor/colleague"<br>○ r: "Working [...] alongside colleagues in research networks"<br>○ r: "Working with the team at the [...] Library" | r=14 |
| \|> Introductions, guides, manuals<br>● r: "Introductions to resources are useful, but it can be hard to know where to find such introductions before you know what you are looking for"<br>● r: "Manual on Gephi"<br>● r: "Repositories help pages and FAQs" | r=4 |

| | |
|---|---|
| ● Penn Library, Lib Guide: Web Archiving for the Arts and Historic Preservation. | |
| \|> Literature<br>● r: "books (obviously)"<br>● r: "Articles by Niels Brügger "<br>● r: "articles about the history of net art and preserving net" | r=3 |
| \|> Information sciences (information studies)<br>● r: "I think having an information science graduate degree is very helpful, although not for specific tools, but more for the general information." | r=1 |
| \|> Providing learner support<br>● r: "Scaffolding technical skill learning" | r=1 |
| \|> Self-learning<br>● r: "Generally looking up YouTube videos on advanced Excel, Power BI, Gephi etc" | r=1 |

### 4.5.2 Data sharing in an institutional or subject repository

Provided with three answer choices, and tick boxes, participants were asked whether they had shared any data they collected or created in an institutional or subject repository. Figure 4.8 offers a representation of participant responses, which shows that more than half of the participants indicated 'No' (61.36%, n=27) followed by 'Yes' (20.45%, n=9), and 8 participants (18.18%) provided no answer. Participants who answered 'Yes' were further asked to name the repository(ies) where their data is stored/shared. 8 participants entered free text responses which were coded into thematic representations.

Table 4.26 offers an overview of such representations which include:

- Repositories (r=4)
- University repository or library (r=3)
- In-vivo representations (r=2)

Mentions of other repositories (r=4) include Zenodo, Institut national de l'audiovisuel, and Dados.gov +. 2 representations are in-vivo and offer alternative interpretations.

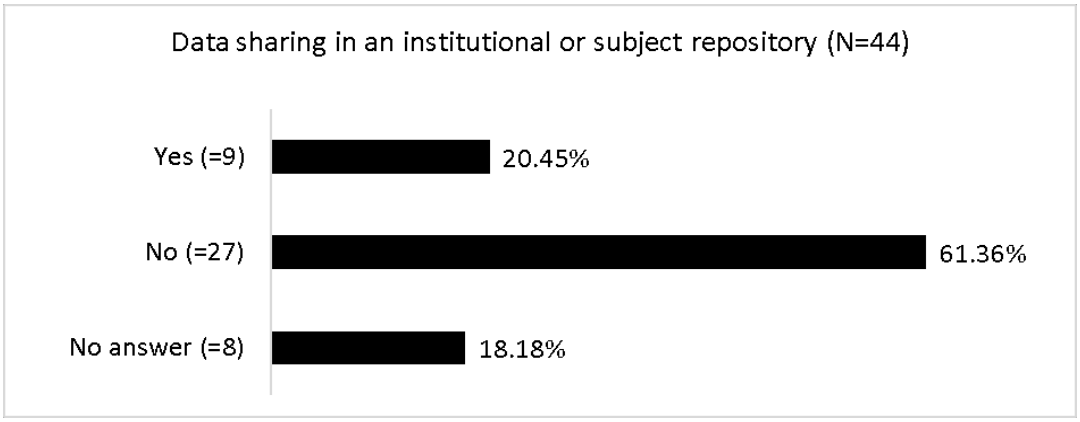


Figure 4.8: Representation of participant responses for whether they shared data in an institutional or library repository (N=44)

Table 4.26: Thematic representation of participant responses for 'Other' repository(ies) used to store/share data (n=8)

| Theme representations or the repository(s) used to store/share data (n=8) | No. of coded representations (R=9) |
|---|---|
| \|> Repositories<br>● Zenodo, https://zenodo.org (r=2)<br>● Institut national de l'audiovisuel, https://www.ina.fr/ (r=1)<br>● Dados.gov +, https://dados.gov.pt/ (r=1) | r=4 |
| \|> University repository / library | r=3 |
| \|> In-vivo representations<br>● r: "some of the data I have collected has been published in articles, books, conference papers and reports and stored on the journal or publisher websites"<br>● r: "Most data I have shared is via web pages on institutional websites, rather than in specific institutional repositories" | r=2 |

## 4.5.3 Final comments

Provided with a comment box, participants were asked if they would like to share any final comments. 10 participants provided free-text comments, of which some merely wrote to express a Thank You. From the comments, 1 participant notes that they are at an early stage of web archiving, and looks forward to learning more, to foster its development. Another participant emphasises the difficulty of archiving the web, yet finds it rewarding, and enjoys learning new skills to figure it out, despite the challenges.

1 participant offers an opinion on further training for web archivists:

- “If WARCnet/IIPC could create course material for web archivists on matters such as how to interpret/use crawl logs, CDX and reports, how to specify crawler settings to scope content in/out, lessons learnt during years of experience, ... that would be very useful. The training materials that have been developed are often on an entry-level, but there is so much more in-depth knowledge available within these networks, it would be wonderful if that could be shared in a structured manner” (WARST Respondent).

1 participant offers an opinion regarding access and interoperability:

- “I am grateful for the [web archive] (ongoing) support for my research. I would be keen for all Web Archives to be publicly and remotely accessible, in the same way that the live web is. I would also [be] keen to see more open and easily accessed interoperability between different countries' web archives” (WARST Respondent).

Several participants indicate that some of the questions in the study were not wholly relevant for them, as outlined below.

- “Basically I am a web archivist and during my [...] research project I was focusing [on a] web archiving project. In this way some aspects of these questions that were focusing on web archives collections as a research subject were just slightly relevant to me” (WARST Respondent).
- "As someone who is primarily focused on web archiving as a means of preserving web art, or artist websites, I found some of these questions not relating to my practice. I have a practical side of the work that I do which rarely needs to practice the skills of the field related to web archiving, because I mainly deal with media files. However, I do keep abreast of the developments in the field. I say this hoping it doesn't skew your data. All the best!” (WARST Respondent).
- “I use web archives for content research rather than data research” (WARST Respondent).

# 5. DISCUSSION

In this study, the participants (N=44) are aged between 18-64 years, indicating that some participants have grown up using the web as a research resource in general, while others have grown up with using more traditional library and archival resources, and had to add the use of web resources to their learning. Nonetheless, in this study, it appears that age has no significant impact on participation in web archive research. In addition, participants identified with residing in North America, Europe, and Asia, and there is an equal representation of participants who identify with being male and female. This is encouraging, as it may provide some indication that gender does not present itself as an obvious barrier in web archive research, in this study at least. To add, the participants (N=44) identify with being at novice, intermediate and experienced levels for working with/using web archives (see Figure 4.3).

In the next section, we organise seven main dimensions for discussion as follows:

- 5.1 Participants - Positions, Backgrounds, and Interests
- 5.2 Pathways to Web Archive Research
- 5.3 Skills and Knowledge Ecologies in Web Archive Research
- 5.4 Challenges with Web Archive Research
- 5.5 Referencing the Archived Web and Data Sharing
- 5.6 Software, Tools, and Methods used in Web Archive Research
- 5.7 Challenges with Legal Deposit, Copyright, and GDPR

## 5.1 Participants - Positions, Backgrounds, and Interests

Regarding the positional background of the participants, we offered two thematic representations being (i) participants who identified with working in a library, archive, or web archive environment (n=30), and (ii) participants who identified as being a scholar, academic, lecturer, post-grad/PhD student, or working in an IT/web design environment (n=14). As mentioned earlier, within this category, 3 participants identified with working in IT or a web design environment outside of academia, but as they are such a small number, we included them in this community, to minimise risks of identification through their responses. Also, to note, there is a much higher representation of participants who identify with being employed in a library, archive, or web archiving environment. With this in mind, we acknowledge that there may be some over-representation by participants from some sectors. However, we feel that this has no effect on the overall aims of the research. Indeed, we consider all opinions to

be valuable when it comes to developing an understanding of web archive research skills, tools, and knowledge. Also, worth mentioning, we initially thought it might be possible to align participants' positions with whether they were creators of web archives, or consumers/users of web archives, but this was not the case. For instance, some respondents in the library, archive, or web archive environment also indicate that they use other web archives as part of their workflows and research. Alternatively, some respondents in the scholar, academic, lecturer, student, or IT/web design environment could also be considered as creators/curators of web archives for research purposes. Thus, the categorisation of participants' positions was not as clear-cut as originally imagined, and we acknowledge that there is some overlap.

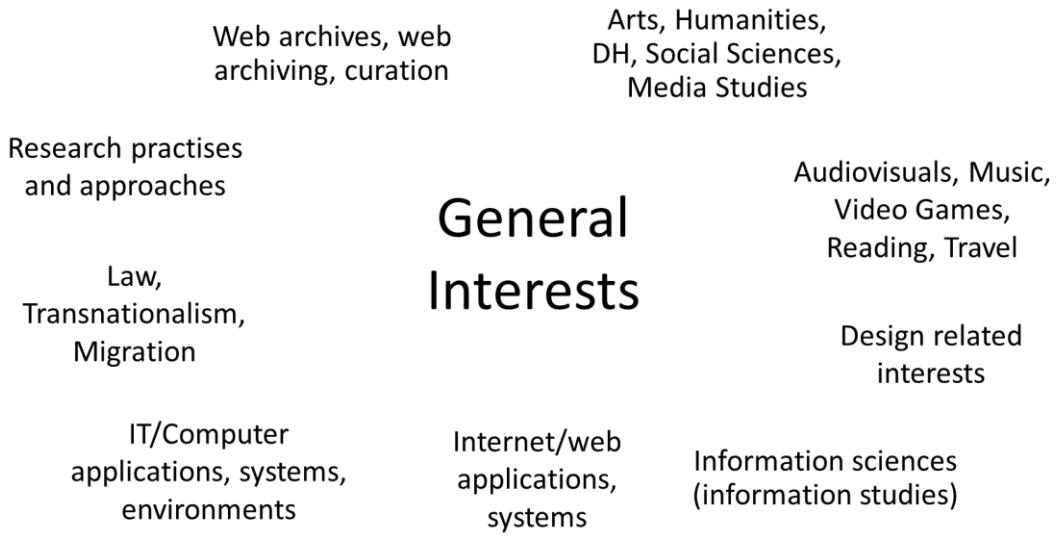


Figure 5.1: WARST participants' interests in general

Overall, the participants' general interests are varied and diverse, transpiring across multiple professional fields, practises, specialisms, and academic disciplines as outlined in Figure. 5.1.

Further to this, broadly based on the participants' interests, backgrounds, experiences, and their relations to web archive research (see Table 4.9), we suggest that the participants in this study identify with one or more of the following subject areas, in alphabetical order (see Figure 5.2).

- Arts, Humanities, DH, Social Sciences, Media Studies
- Business and/or Law
- Data science/analysis, Statistics
- Information sciences (information studies)

- Internet/web applications, systems
- IT/Computer applications, systems, environments
- Use of web archives and archived web content
- Web archives, web archiving, curation

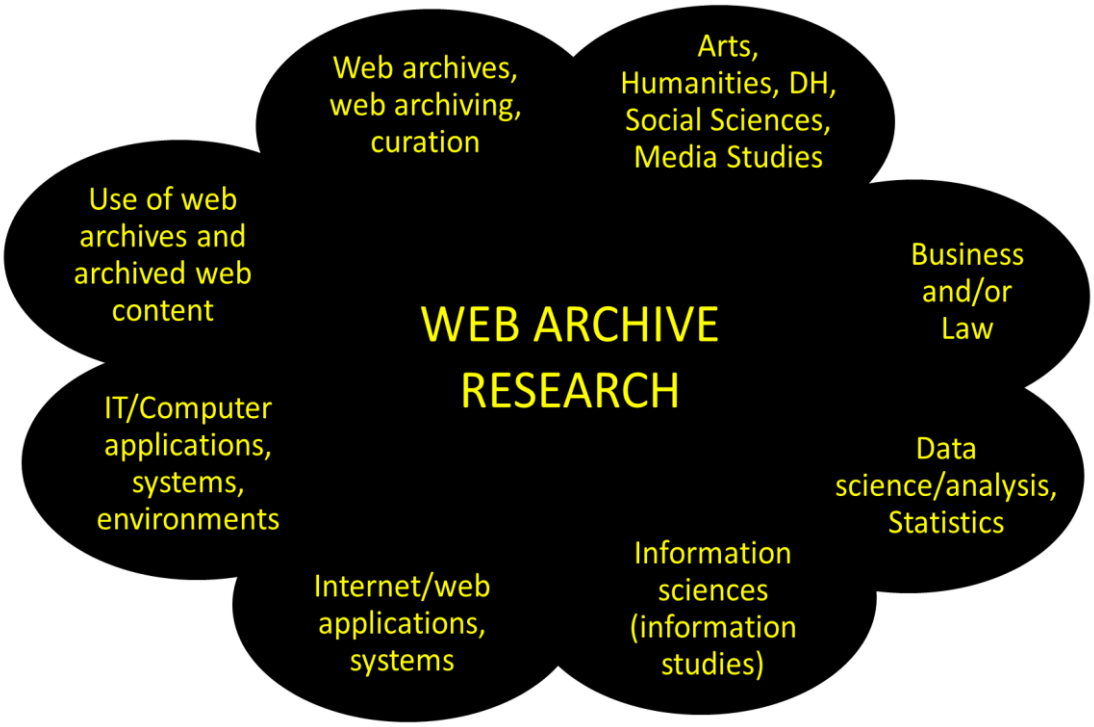


Figure 5.2: In relation to web archive research, the WARST participants identify with one or more of these subject areas

Therefore, it was recognised that it was important to have an interdisciplinary project team conducting this research, due to the diversity of the participants background and interests. The project team includes researchers with a background in humanities, digital humanities, cultural studies, media studies, cultural heritage, library and information science, archival science, computer science, and IT development, and with different skill sets, areas of expertise, and experiences in working with web archives. This was hugely beneficial for contextualising the participants' responses.

## 5.2 Pathways to Web Archive Research

To better understand the pathways which led the participants to curating/using web archives, we pull together two sets of thematic representations from the Results and Analysis and provide them with a label as outlined below.

- Library, Archive, or Web Archive environment - Table 4.10: Thematic representation of responses for reasons which led to curating/using web archives, by participants who identified with Library, Archive, or Web Archive environment (n=28)
- Scholar, Academic, Lecturer, Student, or IT/ Web Design environment - Table 4.11: Thematic representation of responses for reasons which led to using web archives for research, by participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=14)

We further organised the thematic representations from each section, in an alignment as outlined in Table 5.1, bringing the data together as a whole, but with no specific order, or matter of importance.

Table 5.1 offers an overview of the thematic representations for the reasons or pathways which led the participants' involvement in web archive research, in line with participants who identified with a Library, Archive, or Web Archive environment and participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment. We further attempt to connect some commonalities, of which there are a few, while some are open for further interpretation. For example, responses from participants in both communities indicate the use of web archives to find information, literature, and old websites, and show similar concerns about the losses and changes in web content.

Table 5.1: Comparison of thematic representation of participant responses for reasons which led to their involvement in web archive research

| Library, Archive, or Web Archive environment (n=28) | Scholar, Academic, Lecturer, Student, or IT/ Web Design environment (n=14) |
|---|---|
| • Web archives, web archiving, curation<br>• Concerns about the loss/changes of web content<br>• Resource to find information/literature<br>• Business need for a law firm library<br>• r: "Availability during pandemic"<br>• Interests in research aspects/outputs of collections<br>• Digital collection/curation<br>• r: "It is the present and future of archival work."<br>• r: "An adviser taught me how to use it."<br>• r: "My PhD Thesis" | • Resource for conducting research<br>• Concerns about the loss of web content<br>• Resource to find information/old websites<br>• Business need for web content strategy<br>• Ease of access to public web archives<br>• r: "The power of 'raw' internet data to triangulate other data and therefore add to the overall 'scientific' objectivity and credibility of the research"<br>• Richness of data<br>• r: "Web archiving is [a] very important topic, which is not researched enough" |

| | |
|---|---|
| • r: "Internet Archive's Wayback Machine was an early fascination of mine."<br>• r: "The later development of archival tools to capture and catalog websites has been invaluable"<br>• r: "A specific collection for a current [...] senator requires capturing his current website"<br>• Library internship<br>• Subject librarianship | • r: "authoritative source"<br>• r: "Fascination with the centrality of the web in everyday lives and yet its propensity to obsolescence and research oversight"<br>• r: "Wanting [to] make data available" |

## 5.3 Skills and Knowledge Ecologies in Web Archive Research

In a bid for a better understanding of some of the skills and knowledge required for web archive research, we pull together four sets of thematic representations from the Results and Analysis and provide them with a label as outlined below.

- Useful to Have - Table 4.17: Thematic representation of participant responses for 'Other' skills they had before starting their research with web archives which proved useful (n=20)
- Desirable - Table 4.18: Thematic representation of participant responses for other useful skills or knowledge they 'WISH' they had before they started their research in web archives (n=18)
- Acquired - Table 4.19: Thematic representation of participant responses for new skills or knowledge acquired after starting their research in web archives (n=19)
- Also, Useful - Table 4.25: Thematic representation of participant responses for useful resources to further their skills or knowledge in their research with web archives (n=30)

We further organised the themes from each section, in an alignment as outlined in Table 5.2, bringing the data together as a whole, and further organised in descending order of the most common responses. From this, one can see a large array of skills and knowledge that are useful to have, desirable, acquired, and proved to be useful for the participants of this study at least. We outline some of the main representations below.

- Software and tools (r=44)
- Web archives, web archiving, curation (r=21)
- Programming, scripting languages (r=18)
- Digital curation processes/workflows (r=17)

- Data analysis skills (r=13)
- Research methods/approaches (r=11)
- Web design/internet related skills (r=10)
- Information sciences (other than web archiving/curation) (r=9)

Table 5.2 provides a useful interpretation of the skills and knowledge ecologies within the domain of web archive research. The table further signifies the importance of acquiring knowledge and technical and critical skills through training, courses, and workshops, as well as through collaborations and mentorship.

We further suggest that Table 5.2 along with section 4.3.5, challenges encountered when working with web archives, could be used as a starting point for the development of training materials and courses to help overcome some of these challenges. However, we would like to emphasise that in order to develop effective training materials for the skills that are needed to work with web archives, either as a curator, technician or user/researcher, such training would need to be benchmarked in a skills matrix. The Matrix of Digital Curation Knowledge and Competencies developed by Christopher (Cal) Lee provides an excellent template to follow for this future work. It is very hard to develop and provide adequate training without a benchmark to measure against.

Table 5.2: Combined thematic representation of participant responses for skills and knowledge ecologies within web archive research, organised in descending order of the most common responses

| Combined thematic representations for skills and knowledge ecologies within web archive research | Useful to Have (n=20) | Desirable (n=18) | Acquired (n=19) | Also, Useful (n=30) |
|---|---|---|---|---|
| Software and tools (r=44) | r=3 | r=7 | r=18 | r=16 |
| Training, workshops, courses (r=26) | | | | r=26 |
| Web archives, web archiving, curation (r=21) | | | r=21 | |
| Programming, scripting languages (r=18) | r=6 | r=5 | r=7 | |
| Digital curation processes/workflows (r=17) | | | r=17 | |
| Websites, web pages, blogs (r=15) | | | | r=15 |
| Collaborations and mentorship (r=14) | | | | r=14 |
| Data analysis skills (r=13) | r=4 | | r=9 | |
| Research methods/approaches (r=11) | r=8 | | r=3 | |
| Web design/internet related skills (r=10) | r=3 | r=7 | r=3 | |

| | | | | |
|---|---|---|---|---|
| Information sciences (other than web archiving/curation) (r=9) | r=8 | | | r=1 |
| Finding information/services (r=5) | r=3 | r=2 | | |
| Introductions, guides, manuals (r=4) | | | | r=4 |
| Literature (r=3) | | | | r=3 |
| Digital legal deposit (r=2) | | r=1 | r=1 | |
| Languages/translation skills (r=2) | r=2 | | | |
| Managing protected data (r=2) | | r=1 | r=1 | |
| No Skills (r=2) | r=2 | | | |
| Application of metadata (r=1) | | r=1 | | |
| Collaborative skills (r=1) | | r=1 | | |
| Database creation and maintenance (r=1) | | | r=1 | |
| Ethnography (r=1) | | r=1 | | |
| Fair use, copyright, reproduction rights (r=1) | | | r=1 | |
| Glossary of terminology (r=1) | | r=1 | | |
| Graphic design skills (r=1) | r=1 | | | |
| Marketing and public relations (r=1) | | r=1 | | |
| Providing learner support (r=1) | | | | r=1 |
| Self-learning (r=1) | | | | r=1 |
| Skills in usability studies (r=1) | r=1 | | | |
| Social media skills (r=1) | r=1 | | | |
| r: "how indexes are generated, what they contain, and the potential uses they can be put to" | | r=1 | | |
| r: "(hyper)link tracing / retrieval would be useful" | | r=1 | | |
| r: "I really use web archives in a limited capacity and I am not trying to get too fancy." | | r=1 | | |
| r: "All the necessary skills were provided by the [web archive] team" | | r=1 | | |
| r: "Sustainability (long-term availability) of the Internet Archive's Wayback Machine" | | r=1 | | |
| r: "It is hard to list as I would say that I have a fairly advanced knowledge of the computational aspects | | | r=1 | |

| | | | | |
|---|---|---|---|---|
| of working with WARCs at scale, and knew almost nothing starting out." | | | | |
| r: "Most of my digital skills!" | | | r=1 | |

## 5.4 Challenges with Web Archive Research

### 5.4.1 Web archiving, curation, and using web archives for research or other purposes

To better understand the challenges for web archiving and curation, and the use of web archives for research or other purposes, we pull together two sets of thematic representations from the Results and Analysis and provide them with a label as outlined below.

- Library, Archive, or Web Archive environment - Table 4.14: Thematic representation of responses for challenges encountered when working with web archives, by participants who identified with Library, Archive, or Web Archive environment (n=25)
- Scholar, Academic, Lecturer, Student, or IT/Web Design environment - Table 4.15: Thematic representation of responses for challenges encountered when working with web archives, by participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=9)

We further organised the thematic representations from each section, in an alignment as outlined in Table 5.3, bringing the data together as a whole, but with no specific order, or matter of importance. We further attempt to connect some commonalities between the challenges for participants who identified with a Library, Archive, or Web Archive environment and participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment.

Table 5.3 clearly shows multiple challenges which have relevance to each other across both communities of practice. For instance, challenges in capturing dynamic web content may result in archival deficiencies, and incomplete crawls will further translate as inconsistent and incomplete to the end user. Issues for users related to incompleteness in terms of missing image files, and broken links to files such as PDF's or spreadsheets, are also an issue for web archivists. For example, the original link may have been broken on the live site, or changed, during capture. Moreover, Besser (2000) describes the interrelation issues of digital works on the web, in that web pages often incorporate text, images and graphics stored as separate

files, owned by separate organisations, and are often linked to separate servers. This also presents a problem for web archiving initiatives with concerns to “where the boundaries of the work lie” (Besser, 2000).

Dealing with exceptionally large volumes of data is further mentioned as a challenge for respondents from both communities. There are no surprises here. One respondent from the library, archive, and web archiving community notes “Since I am interested in knowing about the entire archive, it means I am interested in multiple Petabytes of data, several million WARC files and Terabytes of index files. The largest barrier has been [the] ability to process this data.” Jackson (2021) offers a meaningful discussion on some of the technical challenges when dealing with big data in the form of domain crawls, and the storage and processing of the same. Challenges in managing and analysing large volumes of data for research purposes are also documented by Truman (2016), Costea (2018), and Healy (2021).

Further challenges arise for web archive users/researchers in the areas of user access, the storage of data transfers from web archives, and the reusability of researcher outputs in the form of derivative data. This is noted as being due to legalities for the archival of web content in the first instance, as well as legalities for providing access to the preserved content, and such legalities vary from country to country. While complications with organising research data that has been extracted from a web archive, under legal deposit/GDPR, have further implications to comprehend. Challenges due to access, sharing and reusability of archived web data, may also be due to interoperability issues across different web archives, as pointed out by one respondent, “I would be keen [...] to see more open and easily accessed interoperability between different countries' web archives.”

In terms of challenges for web archives to organise and provide fully comprehensive documentation and metadata, the following points are noteworthy. First, the provisions of fully comprehensive metadata are problematic when dealing with high volumes of crawled data, as it is time-consuming and labour intensive to provide granular metadata, and it is dependent on the availability of financial resources to do so (Costa, 2021, p. 72; Maemura et al., 2018, p. 1226; Jackson, 2015; Rosenthal, 2015). Consequently, this will affect what the end user will receive in terms of metadata. Thus, it is worthwhile emphasising this aspect to current and potential users. Second, regarding the provisions of comprehensive documentation, challenges often arise due to the legalities which govern acquisition and access which are difficult to describe in pithy, readable documentation for end users, particularly when the end user/researcher community is so diverse, ranging from scholars and academics to members of business and law communities, as well as to members of the general public. There is also the need to consider that end users/researchers may simply not

have the time or energy to invest to acquire a good comprehension of these issues, which may be perceived as a barrier to entry or challenge for engagement with web archives. And on the other side of this web archiving initiatives often do not have the human or financial resources (Costa, 2021) to develop the type of metadata or documentation which would facilitate the diversity of users, who further have different levels of skills and experience. While there are no ready-made solutions for this, there are also indications from this study that there would be some benefit in providing users and potential users with (localised) introductory web archiving training, relative to the web archive being used in a bid to offer more awareness, and thus, more understanding of the scope of the collections vis-à-vis the limitations of archival strategies due to technical challenges, legal constraints, and a lack of resources. In the same way, a traditional archivist might inform a researcher of the limitations of a physical collection directly through a detailed entry in a catalogue, or through query-based communications. It also presents an opportunity for collaboration between web archives and their users to develop documentation in unison, which could eventually be tailored across disciplines and professions.

Challenges in learning new skills are also experienced by respondents from both communities. From the perspective of those working within a web archiving environment, one respondent expresses that the “learning curve was steep”. Another respondent refers to having “Limited technical skills to analyse the WARC-files and the information within them”, and another respondent suggests a challenge in “Learning how to use research tools (from a non-technical user's perspective).” Moreover, for one respondent there is a “Need to learn a lot about what web archives are and the technology that is used to create, curate and maintain them.” From the perspective of a user/researcher, one respondent refers to challenges with “Working with large-scale data and having to acquire new skills (incl learning how to programme with R) in order to perform the necessary analyses.” Another user/researcher suggests “It was difficult to understand the way archives were set up and the tools available to 'talk' to them.” Hence, it seems that both communities would benefit from the provision of training across the full range of activities in the web archiving lifecycle.

Table 5.3: Combined thematic representation of participant responses for challenges encountered in web archive research

| Library, Archive, or Web Archive environment (n=26) | Scholar, Academic, Lecturer, Student, or an IT/Web Design environment (n=9) |
|---|---|
| • Inconsistencies and incompleteness (r=11)<br>• Legalities for acquisition/access (r=8)<br>• Challenges with learning new skills (r=6)<br>• Producing documentation/metadata (r=2)<br>• Volume of data (r=2)<br>• Institutional challenges (r=1)<br>• Technical challenges (r=8)<br>• Financial challenges (r=4)<br>• r: "Having access to the raw data, as a web archivist, is very beneficial" | • Inconsistencies and incompleteness (r=10)<br>• Legalities on access, use, and storage (r=8)<br>• Challenges with learning new skills (=7)<br>• Lack of documentation/metadata (r=2)<br>• Volume of data for research (r=2)<br>• Challenges in an IT/Business/Admin. environment (r=2)<br>• Performance related issues (r=1)<br>• Research methods and approaches (r=5)<br>• r: "One of the big barriers was getting started"<br>• r: "once I wanted to get more involved, who to contact!"<br>• r: "Too many to count!" |

The challenges mentioned above offer strong indications of the need for introductory training for new staff members in a web archiving environment. This is also reflected in the work of Byrne and Rarugal (2019, 2020), who found that 65% of workshop participants (n=26) responded “no” to the question if there was a structured training programme on web archiving at their organisation. Not surprisingly, when these participants were asked ‘how were you trained in web archiving?’, hands-on training was the most popular training method used. As the importance of web archiving grows, so too does the need for training in this field but these responsibilities are falling on web archivists. However, the demands on web archivists' time is always high and it is challenging to find adequate time to develop materials for a structured training programme (Byrne & Rarugal, 2020). Indeed, this is why the IIPC Training Working Group collaborated with the Digital Preservation Coalition (DPC) to develop training materials for beginners. The IIPC established the Training Working Group in October 2017 to “fulfil the vision of making IIPC the world leader for training on web archiving to its members, web archivists and technologists engaged in web archiving” (IIPC, Training Working Group, n.d.). In June 2020, the IIPC Training Working Group launched their first training programme. It comprises slide decks, trainer notes and video case studies that were recorded at the 2019 IIPC Web Archiving Conference (Holownia, 2020). While it seems essential to provide introductory training for incoming web archivists and curators, thereafter, there is a

need to provide a clearly structured plan for consistent, continual training as technologies and approaches change, or upgrade. There is also a need for collaborative efforts to provide more intermediary training, as pointed out below by one respondent.

- "If WARCnet/IIPC could create course material for web archivists on matters such as how to interpret/use crawl logs, CDX and reports, how to specify crawler settings to scope content in/out, lessons learnt during years of experience, ... that would be very useful. The training materials that have been developed are often on an entry-level, but there is so much more in-depth knowledge available within these networks, it would be wonderful if that could be shared in a structured manner" (WARST Respondent).

There are also indications from this study that there would be some value in extending introductory web archiving training to researchers in a bid to offer them more understanding of the limitations of archival strategies due to technical challenges, legal constraints, and a lack of resources. It further indicates that staff in a web archiving environment would also benefit from gaining some understanding and training in the research tools and methods being used by users/researchers to analyse archived web data. Indeed, the study shows that participants from a scholarly or academic environment engage with a diversity of tools and methods and depending on the research question or methodology. Furthermore, such participants also have challenges using archived web for research due to a lack of research methods, theory, and approaches for combining traditional methods with web archive research. Thus, both communities would benefit from collaborative communal training in terms of research approaches and methods for using the archived web, inclusive of demonstrations in tools and software. Indeed, the field would be enriched through the inputs of both communities for developing a better understanding of the research methods and approaches for using web archives, as well as for "Gaining a proper understanding of archived web as a specific type of source and the consequences of these characteristics" for research using the archived web, as pointed out by one respondent.

What also appears evident from various sections of the results, are the number of respondents from both communities who offer indications of the need for collaborations and pathways to develop connections between the creator/curator and the user/researcher. Truman (2016, pp. 3-4) also identifies the need for more communication and collaboration between those who create and steward web archives, and those who use (or might use) a web archive for research. Thus, in this study it is very positive to see acknowledgements of the value of collaborations in practice, and especially how such collaborations benefit both communities in addressing some of the challenges. For example, one respondent notes that "working with specialist archival staff was essential" in order to overcome challenges with

“Closed access, volume, inability to download data, lack of archival context”. Another respondent highlights: “Trying to overcome issues relating to the lack of documentation by establishing close collaborations with curators and IT specialists at the archive”. On the other side of this, one respondent indicates a requirement of their job is to “support researchers who use our web archive collection”, and another expresses an interest in “how to give researchers the best possible access to web archives including tools / API’s etc.”.

Indeed, in this study there are several instances which reflect that some respondents across both communities have a conscious awareness of the importance of such collaborations (e.g., Table 4.25, Table 4.19, Table 4.10). Furthermore, there are indications such collaborations are currently being undertaken to achieve a variety of benefits. For instance, one response mentions a need to work with researchers in order to “promote research use of the archive to lead to more publications citing our archive, with a view to generally increasing usage of the archive + promoting [its] value to our senior stakeholders (particularly funders).” Hence in this instance, supporting researchers enables web archives to develop business cases for more funding leverage, which in turn will develop their services, which will benefit current and future end users in the long term.

Here again, we see the benefits of collaborations between the creators and users of web archives. Winters (2020b) presents a useful demonstration of web archives as “sites of collaboration” to sum up such alliances. Indeed, such collaborations appear to be key for developing current and future practices in the web archive research lifecycle. This was further highlighted in several talks and presentations at the recent IIPC Web Archiving Conference in May 2022 (IIPC, 2022, Conference Abstracts). However, it is worth mentioning that web archiving organisations and institutions may not have the resources to provide the necessary support for researchers. Reasons for this are varied. For example, Brügger (2021) suggests that

> web archives provide the potential for an almost unlimited number of possible forms of researcher interaction, but not all of them can be supported by those archives due to a mix of curatorial, technical, legal, economic and organisational constraints (p. 217).

Such factors may be further influenced by the political and economic climates in a particular country which may not be favourable to funding cultural heritage projects, or indeed may be more favourable to protecting publishers and copyright holders. Other factors are due to a lack of capacity of web archiving organisations to promote the value of web archives to stakeholders (i.e., through user case studies) (cf. Winters, 2020a, p. 170). Here, however, there is a Catch 22 situation, whereby web archiving organisations need resources to assist

researchers to develop user case studies, to demonstrate the value of web archives to attain funding, to provide support to researchers. Thus, for organisations who wish to seek funding to develop web archiving initiatives it is imperative to make a business case for activities in the full web archiving life cycle, inclusive of providing access and support mechanisms for academic researchers and other end users such as journalists or lawyers.

### 5.4.2 Comparison between novice, intermediate and experienced levels

To better understand the challenges for web archiving and curation and the use of web archives, in line with novice, intermediate, and experienced levels, we first use the data from the previous section as follows:

- Library, Archive, or Web Archive environment - Table 4.14: Thematic representation of responses for challenges encountered when working with web archives, by participants who identified with Library, Archive, or Web Archive environment (n=25)
- Scholar, Academic, Lecturer, Student, or IT/ Web Design environment - Table 4.15: Thematic representation of responses for challenges encountered when working with web archives, by participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=9)

We then applied a filter to this data as follows:

- Novice: 0-6 months/6 months - 1 year/ 1-2 years
- Novice/Intermediate : 3-5years
- Intermediate: 5-10 years
- Experienced: 10-15 years/More than 15 years

Table 5.4 offers a breakdown of thematic representations for participant responses for challenges encountered when working with web archives, by participants who identified with working in a Library, Archive or Web Archive environment (n=27), in descending order of most common responses, and in line with novice, intermediate or experienced levels. A full breakdown of this table is available as Appendix C, Table C.1.

Table 5.5 offers an overview of thematic representations for participant responses for challenges encountered when working with web archives, by participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=9), in descending order of most common responses, and in line with novice, intermediate or experienced levels. A full breakdown of this table is available as Appendix C, Table C.2.

The tables below (5.4 & 5.5) highlight the commonalities and differences in challenges encountered by the respondents when working with web archives. By dividing the responses by category of communities of practice and breaking the responses even further by levels of experience in terms of novice, intermediate or experienced, there is no clear trend across different levels of experience. The fact that challenges do not become less with increasing experience highlights the need for training across all levels of experience. Although, in order to develop targeted resources for both introductory and more advanced training, more research would be required to see how challenges shift with increasing experience across communities.

Table 5.4: Combined thematic representations of responses for challenges when working with web archives, by participants who identified with working in a Library, Archive or Web Archive environment (n=27), in line with novice, intermediate or experienced levels

| Theme representations for challenges encountered when working with web archives, by participants who identified with working in a Library, Archive or Web Archive environment (n=27) | Novice 0-2 years | Novice-Intermediate 3-5 years | Intermediate 5-10 years | Experienced 10-15/+15 years |
|---|---|---|---|---|
| Inconsistencies and Incompleteness (r=11) | r=2 | r=3 | r=4 | r=2 |
| Legalities for acquisition/providing access (r=8) | r=3 | r=4 | | r=1 |
| Technical challenges (r=8) | r=2 | r=2 | r=1 | r=3 |
| Challenges with learning new skills (r=6) | r=3 | r=1 | | r=2 |
| Volume of data (r=2) | | r=1 | | r=1 |
| Producing documentation/ metadata (r=2) | r=1 | | r=1 | |
| Financial challenges (r=4) | r=2 | r=1 | | r=1 |
| Institutional challenges (r=1) | | r=1 | | |
| Conceptual challenges (r=1) | | | | r=1 |

Table 5.5: Combined thematic representations of responses for challenges when working with web archives, by participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=9), in line with novice, intermediate or experienced levels

| Theme representations for challenges encountered when working with web archives, by participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=9) | Novice 0-2 years | Novice-Intermediate 3-5 years | Intermediate 5-10 years | Experienced 10-15/ +15 years |
|---|---|---|---|---|
| Inconsistencies and Incompleteness (r=10) | r=1 | r=3 | r=6 | |
| Challenges in an IT/ Business/ Administrative environment (r=2) | r=1 | | r=1 | |
| Challenges with learning new skills (r=6) | | r=3 | r=2 | r=1 |
| Legalities on access, use, and storage (r=8) | | r=3 | r=2 | r=3 |
| Performance related issues (r=1) | r=1 | | | |
| Research methods and approaches (r=5) | | r=3 | r=1 | r=1 |
| Lack of documentation/metadata (r=2) | | | r=1 | r=1 |
| Volume of data for research (r=2) | | | r=1 | r=1 |

## 5.5 Referencing the Archived Web and Data Sharing

### 5.5.1 Referencing styles in general

In terms of referencing practices in general, when using materials other than web archives, participants use a variety of referencing styles such as APA style, MLA style, Harvard style, IEEE style, Chicago style and Turabian style. They further mention using other standards and specifications such as DOI, ISBD, RDA, GOST (ГОСТ) and ISO standards. Some participants note the use of internal or institutional formats, while others suggest that it depends on the journal or publication. Participants also note the use of Zotero, LaTeX or BibTeX. And, for some participants, referencing was not applicable for them. For example, one respondent notes:

“I haven't written academic papers citing web archives (generally, I write policy papers that are about web archiving)”.

### 5.5.2 Referencing archived web materials

Referencing systems are designed to direct a reader to the sources that informed the narrative or conclusions in a body of work, therefore, citation of sources needs to be robust and reliable, inclusive of sources derived from preserved content in a web archive. In this study, just over half of the participants (n=23) indicated that they had ‘No’ challenges citing archived web content, with 16 indicating, ‘Sometime’, and 5 indicating ‘Yes’. So, it seems there is a half-positive perspective, which is encouraging. However, we feel that this area of research might need further investigation as to whether individuals who have no challenges citing archived web content have discovered a useful model which could benefit the community as a whole. It would also be useful to investigate how much disparity there is with the citation practices of individuals with no challenges. For example, a citation may not be a problem for the person citing the content, rather it is a problem for the person using the citation. So, the core function of a citation or reference becomes problematic not only for those creating the citation, but also for those interpreting the citation.

On the other hand, participants who selected ‘Yes’ or ‘Sometimes’ further offered some descriptions of their challenges. Several participants point to a lack of guidelines, standards, or best practices for citing archived web materials, as well as challenges for citing materials from a legal deposit archive, or archives with restrictive access. Also mentioned are challenges that are specific to the URL for archived web content, with one respondent noting: “The standard URL identifier derived from Wayback, while adequate, is unwieldy and not easily read by humans“. For other participants it is simply not easy to cite materials from a web archive. Questions arise here for some participants which include (i) should it be cited like a normal website? (ii) should the source be treated as a normal URL? (iii) should the web archive be acknowledged? (iv) what dates should be used? For instance, one response mentions “what dates should be used (capture date, access date, date of original publication, e.g. a blog post or article).” Another response points to “Ensuring stability of references, even if archive systems change“, while another response offers a solution for referencing archived web content through the incorporation of a PWID URI as a citation aid. A Uniform Resource Identifier for Persistent Web IDentifiers (PWID URI) is a proposed new “web reference standard for archived web references” as a supplement to current citation practises (Zierau et al., 2016, 2018). The fact that there have been research developments in this area also indicates the existence of prior and ongoing challenges for citing materials from a web

archive. Aturban (2019) also describes challenges whereby publicly accessible web archives may be susceptible to link rot if web archive systems change. For example, when a web archiving programme changes their service provider or subscription service, as was the case with the National Library of Ireland Web Archive (NLI Web Archive) who moved their public selective collections from the Internet Memory Foundation to Archive-IT. Respondents also identified challenges for citing materials from a legal deposit archive, or archive with restrictive access, which is problematic for the transparency of the research methods being used. This is further discussed in section 5.7. The challenges described above, certainly warrant more discussion not only between the creators and users of web archives, but also within the wider global arena on the challenges with the citation of evolving born digital and reborn digital media types. Brügger (2016) presents born digital media, as media that has only ever existed in a digital form (such as material on a CD, DVD, the internet, or the web); and reborn digital media, as media that has been collected and preserved and has undergone a change due to this process, such as emulations of computer games or materials in a web archive .

### 5.5.3 Referencing datasets of archived web materials

Less than half of the participants (n=17) responded 'No' to the question of experiencing challenges when citing datasets of archived web content, with 12 participants indicating 'Sometimes', and 7 participants stating 'Yes'. Further to this, participants who answered 'Yes' or 'Sometimes' offered additional descriptions of their challenges. Several participants indicate a lack of guidelines/standards for citing datasets, and some participants indicated that it is not easy to cite datasets in general. Questions are raised in terms of (i) should the web archive be acknowledged in the citation, and (ii) what dates should be used? Another question concerns the amount of data, and what details to include in a dataset citation. Other issues are succinctly summed up by a sample of representations below.

- r:"Amount of detail required is difficult to present in a manner that people can quickly scan and understand ."
- r: "Citing a large corpus that was extracted from [a web archive] with specific parameters, what do you preserve (the actual data, the methods/algorithms/filters/programs) ? - hard for others to redo the research without exact knowledge of the datasets."
- r: "How much to include in relation to describing how the data were collected - depending on context."

Other concerns relate to the data/content reliability of a dataset in terms of its page capture/completeness, and preservation reliability is also mentioned with one respondent

noting: “Derived data sets from web archived data may not be properly preserved”. Ball and Duke (2015) offer a comprehensive overview on the challenges for the citation of datasets in general, which might be used as a starting point to prompt discussion on the challenges for citing datasets with archived web materials.

### 5.5.4 Data sharing

While we were interested in understanding more about the data sharing practices of the participants, it was beyond our scope to examine this in depth in this report. Truter (2021) offers a comprehensive study focused on this area. As part of our report, we queried whether the participants shared any data they collected or created in an institutional or subject repository, and if so, where was it shared. Most participants (n=27) indicated ‘No’ and 9 indicated ‘Yes’. 3 participants note that they share data in a university repository or library. Other respondents mention other repositories such as Zenodo, Institut national de l'audiovisuel and Dados.gov +.

## 5.6 Software, Tools, and Methods Used in Web Archive Research

### 5.6.1 Data collection

To better understand the software and tool ecologies in web archive research, we first pull together 2 sets of thematic representations for tools and methods used for data collection from the Results and Analysis and provide them with a label as outlined below.

- Library, Archive, or Web Archive environment - Table 4.4: Thematic representation of responses for tools and methods used for data collection by participants who identified with Library, Archive, or Web Archive environment (n=30)
- Scholar, Academic, Lecturer, Student, or IT/Web Design environment - Table 4.5: Thematic representation of responses for tools and methods used for data collection by participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=11)

Table 5.6 offers a comparison of thematic representations for the types of tools and methods used by participants for data collection, and Table 5.7 offers a more detailed breakdown of those tools and methods.

The tables reveal that both communities use various capture methods including crawling software, screenshot, screen capture, and screencasting tools, and tools to download data from APIs. Thus, training in these areas would be useful for both communities.

In the library, archive and web archive environment, crawling software which produces data in the standard WARC format predominates. In the scholarly or academic environment, the research question or methodology often influences which tools and methods are chosen, e.g., in cases when data is collected manually for close reading or when only specific parts of a website are scraped. These requirements might explain the greater diversity of tools and methods used by this group of participants.

Web archiving software used for curating and managing web collections is used almost exclusively by participants who identified with working in a library, archive, or web archive environment. This is not surprising, as the effort required for setting up and managing these tools is often too large for personal collections. The software WAIL attempts to reduce these overheads and is used by a participant from a scholarly or academic environment. The use of Archive-It, as a third-party web archiving service is mentioned in both groups, which provides an alternative to managing one's own software for data collection.

Both groups also note the use of tools for replaying web archive content. As the Internet Archive's Wayback Machine is one of the few interfaces that is openly available on the web, it is not surprising that it is used by people from the academic and scholarly community. Respondents from the library and archive environment on the other hand also mention other viewers like OpenWayback and pywb, which are often used for quality control as part of the workflow for selective web archiving. However, it is worth noting that the OpenWayback GitHub currently states that it "is no longer under active development" and suggests that for "high-fidelity replay of web archives, IIPC recommends using Web Recorder's pywb. For those currently hosting instances of OpenWayback, pywb documentation provides a transition guide." Therefore, it might be useful to undertake a study on how web archiving initiatives are coping with the prospects of changing such an important piece of their workflow software.

Changes in web technologies have triggered the development of new tools for data collection. Archiving social media data, for example, typically requires software to download data from a platform-specific API. Tools like Instaloader and Twarc complement traditional crawling software and are mentioned by respondents from both groups. Similarly, different types of browser-based crawling software have been developed to better capture dynamic websites. While respondents from both groups use browser-based crawling software, the diversity is especially marked in the library and archive environment, where six different types of browser-based crawlers are mentioned. Despite these developments, Heritrix with its traditional crawling approach still features frequently in the responses and seems to be the preferred choice for crawling software without browser support.

Table 5.6: Comparison of thematic representation of participant responses for the types of tools and methods used for data collection

| Library, Archive, or Web Archive environment (n=30) | Scholar, Academic, Lecturer, Student, or IT/ Web Design environment (n=11) |
|---|---|
| • Crawling software (r=37)<br>• Curating web archive collections: selection, configuring and scheduling crawls, annotating seeds, performing QA (r=10)<br>• Accessing/replaying archived web data (r=8)<br>• Web archiving subscription services (r=1)<br>• Collecting data from API (r=2)<br>• Managing data (r=5)<br>• Finding source material (r=4)<br>• Screenshot, screen capture (r=2)<br>• Web archiving subscription services (r=1)<br>• Tools with diverse purposes (r=4) (Browser tools, command-line tools, Python scripts/libraries, standard PC tools)<br>• Digital forensics/preservation (r=1)<br>• r: "In house developed web archiving tools"<br>• r: "institutional sources"<br>• r: "text recognition evaluation tools" | • Crawling software (r=7)<br>• Curating web archive collections: selection, configuring and scheduling crawls, annotating seeds, performing QA (r=1)<br>• Accessing/replaying archived web data (r=2)<br>• Web archiving subscription services (r=1)<br>• Collecting data from API (r=2)<br>• Managing data (r=2)<br>• Finding source material (r=6)<br>• Screenshot, screen capture, screencast (r=5)<br>• Web archiving subscription services (r=1)<br>• Tools with diverse purposes (r=4) (Browser tools, Python scripts/libraries, R (Rstudio))<br>• File downloads (r=3)<br>• Web scraping (extracting data from web pages) (r=2)<br>• Audio tools (for interviews) (r=1)<br>• Manual collection for close reading (r=1)<br>• r: "non-English language search words"<br>• r: "direct contact with people who might have the data”<br>• r: "scanning/OCR if the source is hard copy" |

Table 5.7: Comparative breakdown of the tools and methods used for data collection

| Library, Archive, or Web Archive environment (n=30) | Scholar, Academic, Lecturer, Student or being employed in an IT/ Web Design environment (n=11) |
|---|---|
| • Archive-It (r=1)<br>• ArchiveWeb.page (r=4)<br>• Browser tools (r=1)<br>• Browsertrix (r=2)<br>• Brozzler (r=4)<br>• command-line tools (r=1)<br>• Conifer (prior, Webrecorder) (r=9)<br>• CWeb (r=2) | • Archive-It (r=1)<br>• Audio tools (for interviews etc.)<br>• Browser tools (r=2)<br>• Browsertrix (r=1)<br>• Conifer (prior, Webrecorder) (r=2)<br>• Heritrix (r=2)<br>• HTTrack Website Copier (r=1)<br>• Internet Archive, Wayback machine (r=2) |

| | |
|---|---|
| • DSpace (r=1)<br>• Electrolyte (r=3)<br>• Excel, spreadsheet, .csv (=3)<br>• Heritrix (r=11)<br>• HTTrack Website Copier (r=1)<br>• Google Drive (r=1)<br>• Instaloader (r=1)<br>• Internet Archive, Wayback machine (r=3)<br>• Internet, search engines, web search (r=2)<br>• Library catalogues and databases (r=2)<br>• MediaArea tools (r=1)<br>• NetarchiveSuite (r=5)<br>• OpenWayback (r=2)<br>• Python scripts/libraries (r=1)<br>• pywb (r=2)<br>• screen capture tools (in general) (r=1)<br>• snipping tools (in general) (r=1)<br>• Social Feed Manager (r=1)<br>• Umbra (r=1)<br>• W3ACT (r=1)<br>• waybackpy (r=1)<br>• Web crawler (in general) (r=1)<br>• Web Curator Tool (r=1)<br>• Wget (r=1)<br>• r: "selecting material for collection"<br>• r: "In house developed web archiving tools"<br>• r: "institutional sources"<br>• r: " text recognition evaluation tools"<br>• r: "the type of tools that come for standard with a PC" | • Internet, search engines, web search (r=3)<br>• Library catalogues and databases (r=1)<br>• Manual collection for close reading<br>• Manual/scripted file downloads (r=3)<br>• Python scripts/libraries (r=1)<br>• R (Rstudio) (r=1)<br>• screenshot tools/functions (in general) (r=2)<br>• script for screenshot automation (r=1)<br>• SHINE tools - UKWA (r=2)<br>• Snagit (r=1)<br>• Twarc (=1)<br>• Web Archiving Integration Layer (WAIL) (r=1)<br>• Webscraper.io (=1)<br>• web scraping scripts (=1)<br>• Websnapper (r=1)<br>• Wget (r=1)<br>• Zotero (r=1)<br>• Zotfile PlugIn (r=1)<br>• r: "make my own tools to collect data based on [publicly] available API"<br>• r: "non-English language search words"<br>• r: "direct contact with people who might have the data"<br>• r: "scanning/OCR if the source is hard copy" |

## 5.6.2 Data analysis

We then pull together another two sets of thematic representations for tools and methods used for data analysis from the Results and Analysis and provide them with a label as outlined below.

- Library, Archive, or Web Archive environment - Table 4.6: Thematic representation of responses for tools and methods used for data analysis by participants who identified with Library, Archive, or Web Archive environment (n=25)
- Scholar, Academic, Lecturer, Student, or IT/ Web Design environment - Table 4.7: Thematic representation of participant responses for tools and methods used for

data analysis by participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=13)

Table 5.8 offers a comparison of thematic representations for the types of tools and methods used by participants for data collection, and Table 5.9 offers a more detailed breakdown of those tools and methods.

For data analysis, respondents from a library, archive and web archive environment rely heavily on tools for search and information retrieval. While URL-based search is still prevalent in web archives, web archiving institutions have been working to overcome its limitations by complementing it with metadata and full-text search. Today, 72% of web archives around the world offer metadata search, while 63% provide full-text search for all or some of their collections (Costa, 2021, pp. 72–73). Tools like Apache Solr or ElasticSearch as well as relational database technologies, and at a higher level the CDX API are all part of this search infrastructure. While some web archives have also incorporated limited analytical functionality into their user interfaces like network visualisations in the SolrWayback or the trend analysis in the SHINE interface, these services do not feature in the responses from an academic or scholarly environment. Instead, stand-alone tools like Gephi or Nvivo that are not specific to web archive content seem to be used for further analysis. As these tools typically do not support WARC as an input format, further tools like the Archives Unleashed Toolkit or custom scripts and software are used to transform archived web data into formats that are supported by standard analysis software. The fact that tools from the digital humanities and social sciences (Gephi, Voyant Tools, IramuteQ) are also mentioned by some respondents from a library and archive environment, points to an ongoing exchange between these communities.

Table 5.8: Comparison of thematic representation of participant responses for the types of tools and methods used for data analysis

| Library, Archive, or Web Archive environment (n=25) | Scholar, Academic, Lecturer, Student, or IT/ Web Design environment (n=13) |
|---|---|
| ● Collaboration (r=1)<br>● Computer-assisted text analysis (r=2)<br>● Computing infrastructure (r=1)<br>● Data extraction, cleaning, transformation (r=6)<br>● Data management (r=2)<br>● Digital forensics/preservation (r=3)<br>● Distributed processing (r=3) | ● Collaboration (r=1)<br>● Computer-assisted text analysis (r=1)<br>● Qualitative data analysis (r=6)<br>● Data analysis, extraction, cleaning, transformation (r=8)<br>● Programming, scripting languages and computing environments (r=8)<br>● Network analysis (r=3) |

| | |
|---|---|
| ● Evidence analysis (r=1)<br>● Machine learning (r=1)<br>● Metadata, crawl logs (r=3)<br>● Network analysis (r=3)<br>● Programming/scripting languages, computing environments (r=6)<br>● Replay/playback tools (r=2)<br>● Search and information retrieval (r=13)<br>● Statistics (in general) (r=1)<br>● Visualisation (r=4)<br>● Web archive access and analysis (r=1)<br>● Web archiving management (r=1)<br>● r: "lists, notes, tiny pieces of paper"<br>● r: "manual statistics on the report files" from SolrWayback<br>● r: "My work with the web archive involves selecting material, not carrying out research" | ● Other Tools (r=3)<br>● Visualisation (r=1)<br>● r: "mostly my brain"<br>● r: "Conceptual tools (e.g. social semiotics, multimodality) for the [analysis] of complex web objects" |

Table 5.9: Comparative breakdown of the tools and methods used for data analysis

| Library, Archive, or Web Archive environment (n=25) | Scholar, Academic, Lecturer, Student or being employed in an IT/ Web Design environment (n=13) |
|---|---|
| ● Amazon Athena (AWS) (r=1)<br>● Amazon Web Services (r=1)<br>● Apache Hadoop (r=2)<br>● Apache Lucene (r=1)<br>● Apache Parquet (r=1)<br>● Apache Solr (r=1)<br>● Apache Spark (r=1)<br>● Archives Unleashed Toolkit (r=1)<br>● BitCurator (r=1)<br>● Collaboration<br>● CDX queries/files (r=2)<br>● Command-line tools (r=1)<br>● Crawl logs (r=2)<br>● Digiboard (r=1)<br>● DROID (r=1)<br>● ElasticSearch (r=1)<br>● Evidence analysis (=1)<br>● Excel, spreadsheets (r=6)<br>● Gephi (r=3)<br>● GLAM workbench notebooks (r=1)<br>● HeidiSQL/MariaDB (r=1) | ● Archives Unleashed Cloud (r=1)<br>● Archives Unleashed Toolkit (r=1)<br>● Atlas.ti (r=1)<br>● Bash/shell scripting languages (r=3)<br>● Command-line tools (r=1)<br>● Confluence (r=1)<br>● Excel, spreadsheets (r=4)<br>● Gephi (r=3)<br>● Microsoft 365 (r=1)<br>● Nvivo (r=2)<br>● OpenRefine (r=1)<br>● Pattern matching (r=1)<br>● Proprietary tools (r=1)<br>● Perl (r=1)<br>● Python/Python libraries (r=2)<br>● R (r=1)<br>● Regular expressions (r=1)<br>● Voyant tools (r=1)<br>● r: "mostly my brain" |

| | |
|---|---|
| ● IramuteQ (r=1)<br>● Jupyter Notebooks (r=1)<br>● Kibana (r=2)<br>● Lucene (r=1)<br>● MediaArea tools (r=1)<br>● Metadata (r=1)<br>● NutchWax (r=1)<br>● OpenWayback (=1)<br>● Python/Python libraries (r=3)<br>● Pywb (r=1)<br>● R (r=1)<br>● SolrWayback (r=2)<br>● SQL (r=2)<br>● Statistics (in general) (r=1)<br>● statistics on the report files from SolrWayback (r=1)<br>● Tableau (r=2)<br>● TensorFlow (r=1)<br>● Voyant tools (r=1)<br>● r: "Web Archive user interface, faceted functions"<br>● r: "lists, notes, tiny pieces of paper"<br>● r: "manual statistics on the report files" from SolrWayback<br>● r: "My work with the web archive involves selecting material, not carrying out research."<br>● r: "brainstorming with colleagues" | ● r: "Conceptual tools (e.g. social semiotics, multimodality) for the [analysis] of complex web objects"<br>● r: "annotating PDFs with PDFExpert"<br>● r: "Close reading of websites and it's html code"<br>● r: "manual qualitative content analysis"<br>● r: "I usually make my own tools"<br>● r: "visualisation tools for qualitative data" |

### 5.6.3 Other skills, tools, and methods

Other sections of the report also offer insights for various types of skills, tools and methods which are useful for web archive research, as well as insights on areas which would benefit from further discussion and training development. Throughout the findings, we see spreadsheet software being used for the collection, management, and analysis of data by respondents from both communities of practice. We also see the use of spreadsheets as a format for data output. On the other hand, we also see a requirement for training in the use of spreadsheet software, as one respondent notes a “requirement for better knowledge of using spreadsheets in statistical analysis”. Thus, the development of training materials in the use of spreadsheet software, and the management and preservation of spreadsheets as data outputs would be useful from novice to advanced levels for the web archive research community overall.

## 5.7 Challenges with Legal Deposit, Copyright, and GDPR

Across sections of the study, challenges related to legalities, such as legal deposit, copyright and GDPR, are mentioned by respondents from both the web archiving community and the academic community. Moreover, respondents from both groups also discuss challenges for citing archived web content from legal deposit archives, or archives with restrictive access. For example, one respondent notes challenges with citing "historic content" from a restrictive archive, while another respondent notes, "The basic problem is, that if you want to cite to some elements that are in a collection with restricted access, nobody beyond your institution affiliation can check your links." Challenges for copying URLs in legal deposit collections is also pointed out by one respondent in terms of "Copying and pasting a URL from a reading room viewer is not possible as the browsers are locked down." Thus, this becomes problematic for the transparency of the research methods being used.

Several participants who identified with the Library, Archive, or Web Archive environment mention challenges in providing access to archived web collections due to legislation, copyright and GDPR, while another participant mentions challenges in providing access due to embargoes. Another respondent notes that while legal deposit may allow for the collection of websites by a legal deposit institution, it may not effectively deal with the provision of access. Most legal deposit frameworks only allow for institutions to provide access to archived websites onsite. For one respondent this presents a problem as "On-premises access to web archives makes them economically inaccessible." This is a valuable point. Very little attention has been paid to the socio-economic factors which might influence barriers for entry and engagement with web archives, and therefore, is certainly worthy of more targeted research. For those organisations undertaking permission-cleared selective archiving (due to the absence of legal deposit legislation for web archiving), the challenges involved in the acquisition of web content and the provision of access are huge due to the resource-intensive permissions process. Furthermore, while legal deposit may allow for the collection of websites without the need to seek explicit permission for acquisition, in order to make archival copies of websites available offsite, for example, as part of a curated collection, permission is required from the website owner. This presents a challenge, as pointed out by one respondent: "We request that [the owners of] curated websites give us permission to make their material available outside our physical building but many of them simply do not respond."

In the Scholar, Academic, Lecturer, Student, or IT/Web Design environment, several participants discuss challenges in using web archives due to legalities in terms of access to the data, use of the data, storage of the data and the inability to download data from some

web archives. For example, one respondent found challenges working on a transnational collaborative project as due to legal deposit laws in the other country of collaboration, the respondent was unable to view some of the data. As the respondent notes: “I can’t see the actual source code - though my collaborator can - I have to work with statistical data.” Truter (2021) also highlights challenges for researcher/users when it comes to sharing archived web data/materials, due to legal restrictions, including copyright, third-party ownership, privacy policies, and GDPR, which creates challenges for both the use of web archive data and the ability to share the data or make it reusable. Hence, this becomes problematic for researchers in applying for funding, when funders are increasingly stipulating requirements for open access and open science frameworks for research and data outputs (cf. Winters, 2020a, pp. 167–168). It also presents challenges for the development of transnational projects, whereby the researchers involved need access to the same data. This is highlighted by the work of WARCnet Working Group 4, Research Data Management across borders. In addition, when asked about useful skills or knowledge that participants ‘WISH’ they had before they started their research, one respondent notes a requirement for: “Handling protected data (sensitive data and copyright protected data)”. Truter (2021) also suggests that challenges for researchers using web archives may also be due to a lack of training in research data management practices, as well as training for the management and storage of large volumes of protected data. Certainly, further discussion and collaboration is required, to foster developments in the areas of the application of research data management practises within legal deposit frameworks, open science frameworks and web archive research environments.

Finally, in section 4.3.6 we presented findings from participants' responses regarding useful skills or knowledge they had ‘Before’ they started their research with web archives. By filtering further, we examine participant knowledge in how digital legal deposit works and what it is and compare it across both communities of participants in Table 5.10. Table 5.10 offers an overview of participant responses and indicates that the number of participants with no knowledge prior to commencing their research is quite high (9 out of 14) for participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment, and there is also a number of participants (8 out of 30) who identified with Library, Archive, or Web Archive environment. Thus, it seems that introductory training and courses regarding digital legal deposit would be useful for novices from both communities.

Table 5.10: Representation of participant responses for skills and knowledge they had 'Before' they started their research with web archives, in relation to how digital legal deposit works and what it is (N=44)

| How digital legal deposit works and what it is | | | |
|---|---|---|---|
| Library, Archive, or Web Archive environment | (n=30) | Scholar, Academic, Lecturer, Student, or IT/Web Design environment | (n=14) |
| No - I had NO knowledge | =8 | No - I had NO knowledge | =9 |
| Yes - I had a LOT of knowledge | =11 | Yes - I had a LOT of knowledge | =2 |
| Yes - I had SOME knowledge | =11 | Yes - I had SOME knowledge | =3 |

# 6. CONCLUSIONS

This study focuses on individuals around the globe who participate in web archive research, in the context of web archiving, curation, and the use of web archives and archived web content for research or other purposes. We further consider web archive research to be inclusive of the processes and activities described in the Archive-It's web archiving lifecycle model from appraisal, acquisition, and preservation, to replay, access, use and reuse (Bragg & Hannah, 2013). The study sought to identify and document skills, tools, and knowledge required to achieve a broad range of goals within the web archiving lifecycle and explore the challenges for participation in web archive research, and the interludes of such challenges across communities of practice. We suggest that there will always be a need to keep examining the roles of skills, tools, and methods associated with the web archiving lifecycle as long as internet, web and software technologies keep advancing, upgrading, and changing.

In this study, the participants (N=44) are aged between 18-64 years, and identify with residing in North America, Europe, and Asia. Participants identify with being at novice, intermediate and experienced levels for working with, or using web archives, and there is an equal representation of participants who identify with being male and female. This may provide some indication that gender does not present itself as an obvious barrier in web archive research, in this study at least. Regarding the positional background of the participants, we offered two thematic representations being (i) participants who identified with working in a library, archive, or web archive environment (n=30), and (ii) participants who identified as being a scholar, academic, lecturer, post-grad/PhD student, or working in an IT/web design environment (n=14). We initially thought it might be possible to align participants' positions with whether they were creators of web archives, or users of web archives, but this was not the case. For instance, some respondents in the web archiving community also indicate that they are users of various other web archives as part of their workflows and research. Alternatively, some respondents from the scholarly community could also be considered as creators and curators of web archives for research purposes. Thus, the categorisation of participants' positions was not as clear-cut as originally imagined, and we acknowledge that there is some overlap.

Broadly based on the participants' interests, backgrounds, experiences, and their relations to web archive research, we suggest that the participants in this study identify with one or more of the following subject areas, in alphabetical order.

- Arts, Humanities, DH, Social Sciences, Media Studies
- Business and/or Law
- Data science/analysis, Statistics
- Information sciences (other than web archiving/curation)
- Internet/web applications, systems
- IT/Computer applications, systems, environments
- Use of web archives and archived web content
- Web archives, web archiving, curation

## Main Findings and Insights

From the findings, we presented a large array of skills, tools, methods, and knowledge which are required, desirable or useful for the domain of web archive research, across communities of practice. Some of the main representations include:

- Software and tools
- Web archives, web archiving, curation
- Programming, scripting languages
- Digital curation processes/workflows
- Data analysis skills
- Research methods/approaches
- Web design/internet related skills
- Information sciences (other than web archiving/curation)

The study shows several commonalities between participants who identified with working in a library, archive, or web archive environment, and participants who identified as being a scholar, academic, lecturer, student, or participants working in an IT/web design environment. For example, respondents from both communities indicate the use of web archives to find information, literature, and old websites, and show similar concerns about the losses and changes in web content. Dealing with exceptionally large volumes of data is further mentioned as a challenge for respondents from both communities. Also, respondents from both communities indicate the importance of acquiring knowledge and technical and critical skills through training, courses, and workshops, as well as through collaborations and mentorship. What also appears evident from various sections of the results, are the number

of respondents from both communities who offer indications of the need for collaborations and pathways to develop connections between the creator/curator and user/researcher.

In terms of tools and methods, both communities would benefit from training in various capture methods including crawling software, screenshot, screen capture, and screencasting tools, and tools to download data from APIs. There are also indications that the development of training materials in the use of spreadsheet software, and the management and preservation of spreadsheets as data outputs would be useful for novice, intermediate and more advanced levels across the web archive research community as a whole. Furthermore, the study offers indications that users of web archives would benefit from introductory web archiving training, while staff in a web archiving environment would benefit from gaining some understanding and training in the tools and methods being utilised by user/researchers to analyse archived web data. Although, we should point out that the study shows that participants from a scholarly or academic environment engage with a diversity of tools and methods. Moreover, the research question or methodology often influences which tools and methods are chosen, e.g., in cases when data is collected manually for close reading or when only specific parts of a website are scraped. This group of participants also have challenges due to a lack of research methods, theory, and approaches for combining traditional methods with web archive research. Thus, both communities would benefit from collaborative communal training in terms of current research approaches and methods for using the archived web, inclusive of demonstrations in tools and software. In this way, the field would be enriched through the inputs of dialogue by both communities for developing a better understanding of the research methods and approaches for using web archives, as well as for "Gaining a proper understanding of archived web as a specific type of source and the consequences of these characteristics" for research using the archived web, as pointed out by one respondent.

The study presents multiple challenges which have relevance to each other across communities of practice. For example, challenges in capturing dynamic web content may result in archival deficiencies, which may further translate as inconsistent and incomplete to the end user. Issues for users related to incompleteness in terms of missing image files, and broken links to files such as PDF's or spreadsheets, are also an issue for web archivists as the original link may have been broken on the live site, or changed, during capture. Thus, while they are different challenges, they are inextricably linked.

Challenges for end users to access more comprehensive metadata and documentation for web archive collections, are also related to challenges for web archiving initiatives. We note how the provisions of fully comprehensive metadata are problematic when dealing with high

volumes of crawled data, as it is time-consuming and labour intensive and thus, a strain on already limited resources. In addition, a lack of resources, and specialised skill sets will also affect the development of comprehensive documentation, which would facilitate the diversity of users, who further have different levels of skills and experience. There is also a need to consider that academic researchers and other end users such as journalists and lawyers may not have the time or energy to invest to acquire a good comprehension of these issues, and thus, this may be perceived as a barrier to entry or challenge for engagement with web archives. Therefore, there would be some benefit in providing users and potential users with introductory web archiving training, in a localised context relative to the web archive being used in a bid to offer more awareness, and thus, more understanding of the scope of the collections vis-à-vis the limitations of archival strategies due to technical challenges, legal constraints, and a lack of resources. It also presents an opportunity for collaboration between web archives and their users to develop documentation in unison, which could eventually be tailored across disciplines and professions. This would be a significant gain for both communities creating a virtuous circle of creation and end use.

Challenges in learning new skills are also experienced by respondents from both communities. We highlight how both communities would benefit from the provision of collaborative communal training across the full range of activities in the web archiving lifecycle. The study offers an overview of the types of skills and knowledge that web archive creators and web archive users had prior to working with web archives, the skills they developed while working with web archives and the challenges they faced working with this type of resource. We propose that this might be used as a starting point to foster discussions in developing effective training materials for the types of skills and tools that are needed to work with web archives either as a curator, technician, or academic researcher. We further suggest that such training will also need to be benchmarked in a skills matrix, as it is very hard to develop and provide adequate training without a benchmark to measure against. We also find that the challenges experienced by the participants in the study do not become less with increasing experience and highlight the need for training across all levels of experience. Although, we suggest that in order to develop targeted resources for both introductory and more advanced training, further research would be required to see how challenges shift with increasing experience across communities.

Challenges with legalities, such as legal deposit, copyright, and GDPR present other challenges for both the web archiving and researcher/user communities. Respondents from both groups also discuss challenges for citing archived web content from legal deposit archives, or archives with restrictive access. Participants who identified with the web archiving community

mention challenges to provide access to archived web collections due to legislation, copyright, GDPR, and embargoes. Challenges due to low response rates in acquiring permissions from website owners, are also mentioned, for both the capture of sites, as well as to provide access to the archived sites outside of a physical building. Further highlighted is the fact that while legal deposit may allow for the collection of websites by a legal deposit institution, it may not effectively deal with the provision of access. For some institutions, they may only provide access onsite, which "makes them economically inaccessible" as noted by one respondent. This presents an area for more targeted research, as very little attention has been paid to the socio-economic factors which might influence barriers for entry and engagement with web archives.

Participants who identified with the academic community discuss challenges in using web archives due to legalities in terms of access to the data, use of the data, and storage of the data from web archives. Other challenges include handling protected data from a web archive, as well as the inability to download data from some web archives. Challenges working on transnational collaborative projects are also found due to varying legal deposit laws across different countries which affect how the data is accessed, used, and by whom. Moreover, challenges to share data from web archives or make it reusable runs counter to current trends by funders who are increasingly stipulating for open access and open science frameworks for research and data outputs. We suggest that further discussion and collaboration is required, to foster developments in the areas of the application of research data management practises within legal deposit frameworks, open science frameworks, and web archive research environments. As a starting point there would be some benefit in providing introductory training and courses regarding (non-print) digital legal deposit for novices from both communities.

Finally, the study finds positive acknowledgements which reinforces the need and the value of collaborations across communities of practice. The WARST project itself also exemplifies this. The interdisciplinary team of researchers have backgrounds in humanities, digital humanities, cultural studies, media studies, cultural heritage, library and information science, archival science, computer science, and IT development, and have different areas of expertise and experiences in working with web archives which was hugely beneficial when it came to understanding and contextualising the diverse range of participants' responses. The study further highlights how collaborations between web archive creators and users/researchers can benefit both communities in addressing some of the challenges mentioned above. However, we must also acknowledge that web archiving organisations and institutions may not have the resources to provide the necessary support for researchers. Reasons for this are

varied and may be “due to a mix of curatorial, technical, legal, economic and organisational constraints” (Brügger, 2021, p. 217). Such factors may be further influenced by the political and economic climates in a particular country which may not be favourable to funding cultural heritage projects, or indeed may be more favourable to protecting publishers and copyright holders. Other factors are due to a lack of capacity of web archiving organisations to promote the value of web archives to stakeholders (i.e., through user case studies). Indeed, this presents a paradox, whereby web archiving organisations need resources to assist researchers to develop user case studies to demonstrate the value of web archives to attain funding to provide support to researchers. Thus, for organisations who wish to seek funding to develop web archiving initiatives it is imperative to make a business case (from the outset) for activities in the full web archiving life cycle, inclusive of providing access and support mechanisms for academic researchers, and other end users such as journalists or lawyers.

The findings show that due to advances in internet, web, and software technologies, there is a need for the continual evaluation of skills, tools, and methods associated with the full web archiving lifecycle. As one respondent stated web archiving "is the present and future of archival work” but as technologies keep evolving, so too will the challenges. The findings further show the need for creators and users/researchers to keep moving forward as collaborators to guide the next generation of web archive research. As part of this, there will always be a need to keep evaluating skills, tools, and knowledge ecologies in web archive research across communities of practice.

# BIBLIOGRAPHY

The report bibliography is organised in four parts: a list of **References** used in the writing of the report; a list of **Web Archive Providers & Services** mentioned by respondents in the report; a list of resources mentioned by respondents with relevance to **Software, Tools & Methods**; and other **Useful Resources**. The full Bibliography is also available in the project web library in Zotero below.

Zotero | Groups > Skills, Tools, and Knowledge Ecologies in Web Archive Research

We tried to ensure that the URLs provided in the references are (i) captured in a web archive close to the time of access on the live web or (ii) saved by a member of the research team in a web archive close to the time of access on the live web. In case of future link rot, we have documented which archive the URL may be found in, e.g., [URL Memento: Wayback Machine]. An accompanying dataset of bibliographic export files (e.g., BibTex, CSL JSON, CSV, etc.) is also available to download through the WARST Project files, available in Open Science Framework (https://osf.io/vf7gt/). Also as previously mentioned, the glossary for this report is available as a WARCnet Paper.[6]

## References


Adelmann, B., & Franken, L. (2020). Thematic web crawling and scraping as a way to form focussed web archives [Conference abstract]. *Engaging with Web Archives: 'Opportunities, Challenges and Potentialities', (#EWAVirtual), Maynooth University Arts and Humanities Institute, Co. Kildare, Ireland, 21-22 September, 2020* [online]. Retrieved 2020-10-09, from https://zenodo.org/record/4058013. DOI: 10.5281/zenodo.4058013 [URL Memento: Wayback Machine]

Alberts, G., Went, M., & Jansma, R. (2017). Archaeology of the Amsterdam digital city; why digital data are dynamic and should be treated accordingly. *Internet Histories*, *1*(1–2), 146–159. DOI: 10.1080/24701475.2017.1309852

Adobe. (n.d.). [Adobe Web Capture] Capture and archive a website using Adobe Acrobat 9 [Web page]. Adobe Acrobat Library. Retrieved 2022-05-23, from https://acrobatusers.com/tutorials/capture-and-archive-website-using-adobe-acrobat-9. [URL Memento: Wayback Machine]


[6] Healy, S., Byrne, H., Schmid, K., Floody, L., Boté-Vericad, J.-J. (2022) Towards a Glossary for Web Archive Research: Version 1.0. *WARCnet Papers*, September 2022. Aarhus, Denmark: WARCnet (ISSN 2597-0615)

Aleph Archives. (n.d.). UXTR: Universal Links Extractor [Web page]. Aleph Archives. Retrieved 2021-11-29, from http://webarchivingbucket.com/uxtr/doc/. [URL Memento: Wayback Machine]

Alliance of Digital Humanities Organizations. (n.d.). Alliance of Digital Humanities Organizations—ADHO [social media]. ADHO/Facebook. Retrieved 2022-05-07, from https://www.facebook.com/AllianceofDigitalHumanitiesOrganizations/. [URL Memento: archive.today]

Analytical Access to the Domain Dark Archive. (2012+). Analytical Access to the Domain Dark Archive (AADDA) [Blog site]. Analytical Access to the Domain Dark Archive. Retrieved 2019-09-23, from http://domaindarkarchive.blogspot.com. [URL Memento: Wayback Machine]

Anthony, A., Onasoga, K., Ike, D., & Ajayi, O. (2013). Web Archiving: Techniques, Challenges, and Solutions. *International Journal of Management & Information Technology*, 5(3), 598–603. DOI: 10.24297/ijmit.v5i3.760

Archiveteam. (2018+). GeoCities Japan—Archiveteam [Wiki]. Archiveteam/MediaWiki. Retrieved 2022-03-08, from https://wiki.archiveteam.org/index.php/GeoCities_Japan. [URL Memento: Wayback Machine]

Arvidson, A., Persson, K., & Mannerheim, J. (2000, August 13). The Kulturarw3 Project—The Royal Swedish Web Archiw3e—An example of 'complete' collection of web pages. *Proceedings of 66th IFLA Council and General Conference Jerusalem, Israel, 13-18 August 2000.* Retrieved 2021-05-14, from https://archive.ifla.org/IV/ifla66/papers/154-157e.htm. [URL Memento: Wayback Machine]

Association of Internet Researchers. (n.d.). *A*ssociation of Internet Researchers [Website] Association of Internet Researchers. Retrieved 2021-08-11, from https://aoir.org. [URL memento: Wayback Machine]

Aturban, M. (2019, September 10). Where did the archive go? Part 2: National Library of Ireland [Blog post]. Web Science and Digital Libraries Research Group. Retrieved 2019-11-13, from https://ws-dl.blogspot.com/2019/09/2019-09-10-where-did-archive-go-part-2.html. [URL Memento: Wayback Machine]

Bailey, J., Grotke, A., Hanna, K., Hartman, C., McCain, E., Moffatt, C., & Taylor, N. (2014). *Web Archiving in the United States: A 2013 Survey* [NDSA Report]. USA: National Digital Stewardship Alliance. Retrieved 2022-01-27, from https://osf.io/h4e6z/. [URL Memento: archive.today]

Bailey, J., Grotke, A., McCain, E., Moffatt, C., & Taylor, N. (2017). *Web Archiving in the United States: A 2016 Survey* [NDSA Report]. USA: National Digital Stewardship Alliance. Retrieved 2022-01-27, from https://osf.io/hj7rg/. [URL Memento: archive.today]

Ball, A., & Duke, M. (2015). *How to Cite Datasets and Link to Publications: Vol. DCC How-to Guides* (Online/pdf). Edinburgh: Digital Curation Centre. Retrieved 2022-03-11, from

https://www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How_to_Cite_Link.pdf. [URL Memento: archive.today]

Beis, C. A., Harris, K. N., & Shreffler, S. L. (2019). Accessing Web Archives: Integrating an Archive-It Collection into EBSCO Discovery Service. *Journal of Web Librarianship*, 13(3), 246–259. DOI: 10.1080/19322909.2019.1625844

Ben-David, A. (2021). Critical Web Archive Research. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The Past Web: Exploring Web Archives* (pp. 181–188). Cham, Switzerland: Springer.

Besser, H. (2000). Digital Longevity. In Maxine K. Sitts (Ed.), *Handbook for Digital Projects: A Management Tool for Preservation and Access*. Andover, Massachusetts: Northeast Big UK Domain Data for the Arts and Humanities. (n.d.).

Big UK Domain Data for the Arts and Humanities [Website]. Big UK Domain Data for the Arts and Humanities. Retrieved 2021-04-19, from https://buddah.projects.history.ac.uk/. [URL Memento: Wayback Machine]

Bingham, N. (2014). Quality Assurance Paradigms in Web Archiving Pre and Post Legal Deposit. *Alexandria: The Journal of National and International Library and Information Issues*, 25(1-2), 51–68. DOI: 10.7227/ALX.0020

Bingham, N. J., & Byrne, H. (2021). Archival strategies for contemporary collecting in a world of big data: Challenges and opportunities with curating the UK web archive. *Big Data & Society*, 8(1), 1–6. DOI: 10.1177/2053951721990409

Bragg, M., & Hanna, K. (2013). *The Web Archiving Lifecycle Model* [White Paper]. The Archive-It Team, Internet Archive. Retrieved 2021-10-07, from http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf. [URL Memento: Wayback Machine]

Breed, M. (2019). Capturing a Moment: The Practices and Ethics of Social Media Archiving [MA Thesis, The University of North Carolina at Chapel Hill University Libraries]. DOI: 10.17615/P4FF-ZK64

Brügger, N. (2010). Introduction: Web History, an Emerging Field of Study. In N. Brügger (Ed.), *Web History* (pp. 01–26). New York: Peter Lang.

Brügger, N. (Ed.). (2010). *Web History*. New York: Peter Lang.

Brügger, N. (2016). Digital Humanities in the 21st Century: Digital Material as a Driving Force. *Digital Humanities Quarterly*, 10(2). Retrieved 2018-11-09, from http://www.digitalhumanities.org/dhq/vol/10/3/000256/000256.html. [URL Memento: Wayback Machine]

Brügger, N. (Ed.). (2017). *Web 25: Histories from the first 25 Years of the World Wide Web* (Vol. 112). New York: Peter Lang.

Brügger, N. (2018). *The Archived Web: Doing History in the Digital Age*. Massachusetts, London: The MIT Press.

Brügger, N. (2020). Welcome to WARCnet. *WARCnet Papers*. Aarhus, Denmark: WARCnet. Retrieved 2021-01-27, from

https://cc.au.dk/fileadmin/user_upload/WARCnet/1.Bru__gger_Welcome_to_WARCnet.pdf [URL Memento: Wayback Machine]

Brügger, N. (2021). The Need for Research Infrastructures for the Study of Web Archives. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The Past Web: Exploring Web Archives* (pp. 217–224). Cham, Switzerland: Springer. DOI: 10.1007/978-3-030-63291-5_17

Brügger, N., & Finnemann, N. O. (2013). The Web and Digital Humanities: Theoretical and Methodological Concerns. *Journal of Broadcasting & Electronic Media*, 57(1), 66–80. DOI: 10.1080/08838151.2012.761699

Brügger, N., & Laursen, D. (Eds.). (2019). *The Historical Web and Digital Humanities: The Case of National Web Domains*. London & New York: Routledge.

Brügger, N., & Milligan, I. (Eds.). (2019). *The SAGE Handbook of Web History*. London: SAGE Publications.

Brügger, N., & Schroeder, R. (Eds.). (2017). *The Web as History: Using Web Archives to Understand the Past and the Present* London: UCL Press. DOI: 10.14324/111.9781911307563; Online/pdf. [URL Memento: Wayback Machine]

Byrne, H. (2020, September 10). Launching the UK Web Archive 2020 Annual Domain Crawl [Blog post]. *UK Web Archive Blog*. Retrieved 2021-05-30, from https://blogs.bl.uk/webarchive/2020/09/launching-the-uk-web-archive-2020-annusal-domain-crawl.html. [URL Memento: Wayback Machine]

Byrne, H., & Rarugal, C. (2019, June 6). Workshop: Reflecting on how we train new starters in web archiving. *International Internet Preservation Coalition General Assembly and Web Archiving Conference, Zagreb, Croatia, 6-7 June 2019*. Retrieved 2022-02-02, from https://digital.library.unt.edu/ark:/67531/metadc1609017/. [URL Memento: archive today]

Byrne, H., & Rarugal, C. (2020, May 24). Reflecting on how we train new starters in web archiving [Blog post]. *IIPC Blog*. Retrieved 2020-05-24, from https://netpreserveblog.wordpress.com/2020/05/24/reflecting-on-how-we-train-new-starters-in-web-archiving/. [URL Memento: Wayback Machine]

Cho, J., & Garcia-Molina, H. (2000). The Evolution of the Web and Implications for an Incremental Crawler. *Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, Cairo, Egypt, 2000*. Retrieved 2021-11-23, from http://www.vldb.org/conf/2000/P200.pdf. [URL Memento: Wayback Machine]

Chudoba, B. (n.d.). How much time are respondents willing to spend on your survey? [Web page]. SurveyMonkey. Retrieved 2018-05-17, from https://www.surveymonkey.com/curiosity/survey_completion_times/. [URL Memento: Wayback Machine]

CCDSS-DAI, Consultative Committee for Space Data Systems (CCSDS), Data Archive Interoperability (DAI) Working Group. (2021, September 1). Data Archive Interoperability Working Group Presentations | CCSDS.org [Conference keynote presentation]. *Engaging with Web Archives for Digital Humanities, 2021, Maynooth*

*University Arts and Humanities Institute, Co. Kildare, Ireland*. Retrieved 2020-09-22, from https://public.ccsds.org/outreach/DAIVideos.aspx. [URL Memento: archive.today]

CollEx - Persée. (n.d.). ResPaDon < CollEx—Persée [Web page]. *CollEx - Persée*. Retrieved 2021-10-15, from https://www.collexpersee.eu/projet/respadon/. [URL Memento: Wayback Machine]

CoolTool. (2017). 6-10 Minutes Is the Ideal Survey Length [Blog post]. *CoolTool Blog*. Retrieved from https://web.archive.org/web/20201111232524/https://cooltool.com/blog/6-10-minutes-is-the-ideal-survey-length [Web archive: Wayback Machine; source URL: https://cooltool.com/blog/6-10-minutes-is-the-ideal-survey-length; Timestamp: 2020-11-11 23:25:24]

Costa, M. (2021). Full-Text and URL Search Over Web Archives. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The Past Web: Exploring Web Archives* (pp. 71–84). Cham, Switzerland: Springer. DOI: 10.1007/978-3-030-63291-5_7

Costa, M., & Silva, M. J. (2010). Understanding the information needs of web archive users. In J. Masanès, A. Rauber, & M. Spaniol (Eds.), *Proceedings of the 10th International Web Archiving Workshop (IWAW '10), Vienna, Austria, September 22-23, 2010*, 9–16. IWAW. Retrieved from https://web.archive.org/web/20110723173820/http://www.iwaw.net/10/IWAW2010.pdf [Web archive: Wayback Machine; source URL: http://www.iwaw.net/10/IWAW2010.pdf; Timestamp: 2011-07-23 17:38:20]

Costea, M.-D. (2018). *Report on the Scholarly Use of Web Archives* [Report]. Aarhus: NetLab. Retrieved 2019-08-30, from http://netlab.dk/wp-content/uploads/2018/02/Costea_Report_on_the_Scholarly_Use_of_Web_Archives.pdf [URL Memento: Wayback Machine]

Day, M. (2006). The Long-Term Preservation of Web Content. In J. Masanès (Ed.), *Web Archiving* (pp. 177–199). Berlin, Heidelberg: Springer-Verlag.

De Haan, T. (2018, November 29). Bit by bit, byte by byte: Web archaeology going strong in the Netherlands! [Web page]. *DPC Blog.* Retrieved 2021-10-11, from https://www.dpconline.org/blog/wdpd/bit-by-bit-byte-by-byte. [URL Memento: Wayback Machine]

De Haan, T., Jansma, R., & Vogel, P. (2017). *DIY Handboek voor Webarcheologie* [Guide]. Amsterdam Museum. Retrieved 2021-09-16, https://hart.amsterdam/image/2017/11/17/20171116_freeze_diy_handboek.pdf. [URL Memento: Wayback Machine]

Denev, D., Mazeika, A., Spaniol, M., & Weikum, G. (2009). SHARC: Framework for Quality-conscious Web Archiving. *Proceedings of the VLDB Endowment*, 2, 586–597. DOI: 10.14778/1687627.1687694

Denev, D., Mazeika, A., Spaniol, M., & Weikum, G. (2011). The SHARC Framework for Data Quality in Web Archiving. *The VLDB Journal*, 20(2), 183–207. DOI: 10.1007/s00778-011-0219-9

Dougherty, M. (2007). Archiving the Web: Collection, Documentation, Display, and Shifting Knowledge Production Paradigms [PhD Dissertation, University of Washington]. ProQuest One Academic. ProQuest document ID: 304794229

Dougherty, M., Meyer, E. T., Madsen, C., McCarthy, van den Heuvel, C., Thomas, A., & Wyatt, S. (2010). *Researcher Engagement with Web Archives: State of the Art* (Joint Information Systems Committee Report, August 2010). London: Joint Information Systems Committee (JISC). Retrieved 2020-07-31, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1714997. [URL Memento: Wayback Machine]

Dupont, H. (1999). Legal Deposit in Denmark—The New Law and Electronic Products. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 9(2), 244–251. DOI: 10.18352/lq.7539

Eltgroth, D. (2009). Best Evidence and the Wayback Machine: Toward a Workable Authentication Standard for Archived Internet Evidence. *Fordham Law Review*, 78(1), 181. Retrieved 2021-04-28, from https://ir.lawnet.fordham.edu/flr/vol78/iss1/5. [URL Memento: Wayback Machine]

European Commission. (n.d.). Data protection [Web page]. European Commission. Retrieved 2021-11-30, from https://ec.europa.eu/info/law/law-topic/data-protection_en. [URL Memento: Wayback Machine]

EWA Conference. (2019+). #EWA Conference (@EWAConf) / Twitter [social media]. Twitter. Retrieved 2022-05-07, from https://twitter.com/EWAConf. [URL Memento: archive.today]

Farrell, M., McCain, E., Praetzellis, M., Thomas, G., & Walker, P. (2018). *Web Archiving in the United States: A 2017 Survey* [NDSA Report]. USA: National Digital Stewardship Alliance (NDSA). Retrieved 2021-02-25, from https://osf.io/r5pqk/. DOI 10.17605/OSF.IO/R5PQK [URL Memento: archive.today]

Fetterly, D., Manasse, M., Najork, M., & Wiener, J. (2003). A large-scale study of the evolution of web pages. *Proceedings of the 12th International Conference on World Wide Web (WWW '03)*, 669–678. DOI: 10.1145/775152.775246

Germain, C. A. (2000). URLs: Uniform resource locators or unreliable reliable resource locators? *College and Research Libraries*, 61(4), 359–365. College & Research Libraries. DOI: 10.5860/crl.61.4.359 [URL Memento: archive.today]

Gomes, D., & Costa, M. (2014). The Importance of Web Archives for Humanities. *International Journal of Humanities & Arts Computing: A Journal of Digital Humanities*, 8(1), 106–123. DOI: 10.3366/ijhac.2014.0122

Gomes, D., Demidova, E., Winters, J., & Risse, T. (Eds.). (2021). *The Past Web: Exploring Web Archives* (Hardback). Cham, Switzerland: Springer. [Preprint:

https://sobre.arquivo.pt/wp-content/uploads/The-Past-Web_-exploring-web-archives-preprint.pdf]

Gomes, D., Miranda, J., & Costa, M. (2011). A Survey on Web Archiving Initiatives. In S. Gradmann, F. Borri, C. Meghini, & H. Schuldt (Eds.), *Research and Advanced Technology for Digital Libraries.* (Vol. 6966). DOI: 10.1007/978-3-642-24469-8_41

Gorsky, M. (2015). Into the Dark Domain: The UK Web Archive as a Source for the Contemporary History of Public Health. *Social History of Medicine*, 28(3), 596–616. DOI: 10.1093/shm/hkv028

Gottsegen, G. (2018, October 2). GeoCities dies in March 2019, and with it a piece of internet history [News article]. CNET. Retrieved 2021-09-05, from https://www.cnet.com/tech/services-and-software/geocities-dies-in-march-2019-and-with-it-a-piece-of-internet-history/. [URL Memento: Wayback Machine]

Graham, P. M. (2017). Guest Editorial: Reflections on the Ethics of Web Archiving. *Journal of Archival Organization*, 14(3–4), 103–110. DOI: 10.1080/15332748.2018.1517589

Government of Ireland. Copyright and Other Intellectual Property Law Provisions Act 2019, https://www.irishstatutebook.ie/eli/2019/act/19/enacted/en/print.html. [URL Memento: Wayback Machine]

Grotke, A., & Jones, G. (2010). DigiBoard: A Tool to Streamline Complex Web Archiving Activities at the Library of Congress. In J. Masanès, A. Rauber, & M. Spaniol (Eds.), *Proceedings of the 10th International Web Archiving Workshop (IWAW '10), Vienna, Austria, September 22-23, 2010,* 17–23. Retrieved from https://web.archive.org/web/20110723173820/http://www.iwaw.net/10/IWAW2010.pdf [Web archive: Wayback Machine; source URL: http://www.iwaw.net/10/IWAW2010.pdf; Timestamp: 2011-07-23 17:38:20]

Harter, S. P., & Kim, H. J. (1996). Electronic journals and scholarly communication: A citation and reference study. *Information Research*, 2(1). Retrieved 2021-02-26, from http://informationr.net/ir/2-1/paper9a.html. [URL Memento: Wayback Machine]

Healy, S. (2019, November 30). Web archives as resources to find archived treasures. MU Library Treasures. Retrieved 2022-02-12, from https://mulibrarytreasures.wordpress.com/2019/11/30/web-archives-as-resources-to-find-archived-treasures/. [URL Memento: archive.today]

Healy, S. (2021) Awareness and Engagement with Web Archives in Irish Academic Institutions. *EdTech Winter Online Conference 2021 Paradigm Shift : Reflection, Resilience and Renewal in Digital Education, 14-15 January, 2022.* Irish Learning Technology Association. Retrieved 2021-08-17, from https://edtech2021.exordo.com/programme/presentation/95. [URL Memento: Wayback Machine]

Healy, S., Byrne, H., Schmid, K., (2022) Zotero | Groups > Skills, Tools, and Knowledge Ecologies in Web Archive Research. Zotero, https://www.zotero.org/groups/4669886/skills_tools_and_knowledge_ecologies_in_web_archive_research [export files available in OSF: https://osf.io/vf7gt/]

Healy, S., Byrne, H., Schmid, K., Floody, L., Boté-Vericad, J.-J. (2022) Towards a Glossary for Web Archive Research: Version 1.0. *WARCnet Papers*, September 2022. Aarhus, Denmark: WARCnet (ISSN 2597-0615)

Healy, S., Byrne, H., Schmid, K., Floody, L., Boté-Vericad, J.-J. (2021+) Zotero | Groups > Towards a Glossary for Web Archive Research. Zotero, https://www.zotero.org/groups/4380600/towards_a_glossary_for_web_archive_research. [export files available in OSF: https://osf.io/vf7gt/]

Hockx-Yu, H. (2011). The Past Issue of the Web. *Proceedings of the 3rd International Web Science Conference (WebSci '11)*, 1-8 (Article 12). DOI: 10.1145/2527031.2527050

Hockx-Yu, H. (2014). Access and Scholarly Use of Web Archives. *Alexandria*, 25(11), 113–127. DOI: 10.7227/ALX.0023

Holownia, O. (2020, June 15). Launching IIPC training programme [Blog post]. *IIPC Blog*. Retrieved 2022-01-10, from https://netpreserveblog.wordpress.com/2020/06/15/launching-iipc-training-programme. [URL Memento: Wayback Machine]

Holzmann, H., & Nejdl, W. (2021). A Holistic View on Web Archives. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The Past Web: Exploring Web Archives* (pp. 85–99). Cham, Switzerland: Springer. DOI: 10.1007/978-3-030-63291-5_8

Huurdeman, H. C., & Kamps, J. (2017). A Collaborative Approach to Research Data Management in a Web Archive Context. In *Research Data Management—A European Perspective* (pp. 55–78). Berlin/Boston: De Gruyter Saur. DOI: 10.1515/9783110365634-005

Huc-Hepher, S., & Wells, N. (2021). Exploring Online Diasporas: London's French and Latin American Communities in the UK Web Archive. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The Past Web: Exploring Web Archives* (pp. 189–201). Cham, Switzerland: Springer. DOI: 10.1007/978-3-030-63291-5_15

International Federation of Library Associations and Institutions. (n.d.). DIGLIB—Digital Libraries Research Mailing List [Web page]. IFLA Mailing Lists Service. Retrieved 2022-05-07, from https://mail.iflalists.org/wws/info/diglib. [URL Memento: Wayback Machine]

International Internet Preservation Consortium. (n.d.). International Internet Preservation Consortium [Website]. International Internet Preservation Consortium. Retrieved 2021-04-23, from https://netpreserve.org/. [URL Memento: Wayback Machine]

International Internet Preservation Consortium. (n.d.). Legal deposit | IIPC [Web page]. International Internet Preservation Consortium. Retrieved 2021-11-03, from https://netpreserve.org/web-archiving/legal-deposit/. [URL Memento: Wayback Machine]

International Internet Preservation Consortium. (n.d.). Tools & software | IIPC [Web page]. *International Internet Preservation Consortium*. Retrieved 2021-01-01, from http://netpreserve.org/web-archiving/tools-and-software/. [URL Memento: Wayback Machine]

International Internet Preservation Consortium. Awesome Web Archiving | IIPC [GitHub]. International Internet Preservation Consortium (iipc/awesome-web-archiving). Retrieved 2021-02-23, from https://github.com/iipc/awesome-web-archiving. [URL Memento: Wayback Machine]

International Internet Preservation Consortium. (n.d.). Training materials | IIPC [Web page]. *International Internet Preservation Consortium*. Retrieved 2021-07-21, from http://netpreserve.org/web-archiving/training-materials/. [URL Memento: Wayback Machine]

International Internet Preservation Consortium. (n.d.). Training Working Group | IIPC [Web page/pdf]. International Internet Preservation Consortium. Retrieved 2022-03-25, from https://netpreserve.org/about-us/working-groups/training-working-group/. [URL Memento: Wayback Machine]

International Internet Preservation Consortium. (n.d.). Web Archiving: Why archive the web? | IIPC [Web page]. International Internet Preservation Consortium. Retrieved 2021-12-29, from https://netpreserve.org/web-archiving/. [URL Memento: Wayback Machine]

Jackson, A., Lin, J., Milligan, I., & Ruest, N. (2016). Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities. *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, 103–106. DOI: 10.1145/2910896.2910912

Jackson, A. (2015, November 20). The Provenance of Web Archives [Blog post]. UK Web Archive Blog. Retrieved 2022-02-24, from https://britishlibrary.typepad.co.uk/webarchive/2015/11/. [URL Memento: Wayback Machine]

Jackson, A. (2022, January 6). UKWA 2021 Technical update [Blog post]. UK Web Archive Blog. Retrieved 2022-01-07, from https://blogs.bl.uk/webarchive/2022/01/ukwa-2021-technical-update.html. [URL Memento: Wayback Machine]

Jacobsen, G. (2008). Web Archiving: Issues and Problems in Collection Building and Access. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 18(3–4), 366–376. DOI: 10.18352/lq.7936 [URL Memento: Wayback Machine]

Jansma, R. (2020). Scoops and Brushes for Software Archaeology: Metadata Dating [Vrije Universiteit Amsterdam; Universiteit van Amsterdam]. Retrieved 2021-04-22, from https://jansma.io/Papers/Scoops_and_Brushes_for_Software_Archaeology_-_Metadata_Dating.pdf. [URL Memento: Wayback Machine]

Jatowt, A., Kawai, Y., Ohshima, H., & Tanaka, K. (2008). What can history tell us? Towards different models of interaction with document histories. *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia*, 5–14. DOI: 10.1145/1379092.1379098

JISC Online Surveys. (n.d.). JISC Online Surveys [Website]. Joint Information Systems Committee Online Surveys. Retrieved 2021-04-17, from https://www.onlinesurveys.ac.uk/. [URL Memento: Wayback Machine]

Ken Web Archiving. (n.d.). Ken | A Better Way to Control Enterprise Data [Website]. Ken Web Archiving. https://ken-webarchiving.com. [URL Memento: Wayback Machine]

Kitchens, J. D., & Mosley, P. A. (2000). Error 404: Or, what is the shelf-life of printed Internet guides? *Library Collections, Acquisitions, & Technical Services*, 24(4), 467–478. Elsevier, ScienceDirect. DOI: 10.1016/S1464-9055(00)00178-0

Koehler, W. (1999). Digital Libraries and World Wide Web Sites and Page Persistence. *Information Research*, 4(4). Information Research. Retrieved 2021-02-11, from http://www.informationr.net/ir/4-4/paper60.html. [URL Memento: Wayback Machine]

Koerbin, P. (2021). National Web Archiving in Australia: Representing the Comprehensive. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The Past Web: Exploring Web Archives* (pp. 23–32). Cham, Switzerland: Springer. DOI: 10.1007/978-3-030-63291-5_3

Kurzmeier, M. (2021). Political Expression in Web Defacements [PhD Dissertation, Maynooth University]. DOI: 10.5281/zenodo.6308125

Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the web. *Nature*, *400*(6740), 107–107. DOI: 10.1038/21987 [URL Memento: Wayback Machine]

Lee, C. (2017). Matrix of Digital Curation Knowledge and Competencies (Overview)—DigCCurr Project (Version 17, Online/web). School of Information and Library Science, University of North Carolina at Chapel Hill. Retrieved 2022-01-20, from https://ils.unc.edu/digccurr/digccurr-matrix.html. [URL Memento: Wayback Machine]

Leetaru, K. (2017, January 13). Why Aren't We Doing More with Our Web Archives? *Forbes*, *AI & Big Data*. Retrieved 2021-01-24, from https://www.forbes.com/sites/kalevleetaru/2017/01/13/why-arent-we-doing-more-with-our-web-archives/?sh=155f433c498a. [URL Memento: archive.today]

Library of Congress, Public Affairs Office. (1998, October 13). Alexa Internet Donates Archive of the World Wide Web to Library of Congress: First Large-Scale Digital Donation Ensures Preservation of Digital Cultural Artifacts [web version]. *News from the Library of Congress*, *PR 98-167*. Retrieved from https://web.archive.org/web/20030423175610/http://www.loc.gov/today/pr/1998/98-167.html. [Web archive: Wayback Machine; source URL: http://www.loc.gov/today/pr/1998/98-167.html; Timestamp: 2003-04-23 17:56:10]

Lyman, P. (2002). Archiving the World Wide Web. In Council on Library and Information Resources & Library of Congress (Eds.), *Building a National Strategy for Preservation: Issues in Digital Media Archiving* (Online/pdf, pp. 38–51). USA: Council on Library and Information Resources and the Library of Congress. Retrieved 2021-03-29, from https://www.clir.org/wp-content/uploads/sites/6/pub106.pdf. [URL Memento: Wayback Machine]

MAXQDA Blog. (2021, June 21). MAXQDA Tip of the month: In-vivo coding. *MAXQDA Blog*. Retrieved 2021-06-21, from https://www.maxqda.com/blogpost/tip-of-the-month-in-vivo-coding-out-of-the-document. [URL Memento: Wayback Machine]

Mackinnon, K. (2021). Ethical Approaches to Youth Data in Historical Web Archives (Dispatch). *Studies in Social Justice*, 15(3), 442–449. DOI: 10.26522/ssj.v15i3.2541

Mackinnon, K. (2022). The death of GeoCities: Seeking destruction and platform eulogies in Web archives. *Internet Histories*, 0(0), 1–16. DOI: 10.1080/24701475.2022.2051331

Maemura, E. (2018). What's cached is prologue: Reviewing recent web archives research towards supporting scholarly use. *Proceedings of the Association for Information Science and Technology*, 55(1), 327–336. DOI: 10.1002/pra2.2018.14505501036

Maemura, E. (2022). Towards an Infrastructural Description of Archived Web Data. *WARCnet Papers*. Aarhus, Denmark: WARCnet. Retrieved 2022-05-09, from https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Maemura_Towards_an_Infrastructural_Description.pdf. [URL Memento: Wayback Machine]

Masanès, J. (2005). Web Archiving Methods and Approaches: A Comparative Study. *Library Trends*, 54(1), 72–90. DOI: 10.1353/lib.2006.0005

Masanès, J. (2006). Web Archiving: Issues and Methods. In J. Masanès (Ed.), *Web Archiving* (pp. 1–53). Berlin, Heidelberg: Springer-Verlag.

MAXQDA. (n.d.). MAXQDA | All-In-One Qualitative & Mixed Methods Data Analysis Tool [Website]. MAXQDA. Retrieved 2021-08-18, from https://www.maxqda.com/. [URL Memento: Wayback Machine]

Maynooth University. (2016). Maynooth University Research Integrity Policy. Maynooth University. Retrieved 2017-12-02, from https://www.maynoothuniversity.ie/sites/default/files/assets/document/MU%20Research%20Integrity%20%20Policy%20September%202016%20_2.pdf. [URL Memento: Wayback Machine]

Maynooth University. (2019). Maynooth University Online Surveys (formerly Bristol Online Survey) User Policy. Maynooth University. Retrieved 2022-02-15, from https://www.maynoothuniversity.ie/sites/default/files/assets/document/Maynooth%20University%20OnlineSurveys%20User%20Policy%20FINAL.pdf. [URL Memento: archive.today]

Maynooth University. (2020). Maynooth University Research Ethics Policy (Updated March 2020). Maynooth University. Retrieved 2020-11-10, from https://www.maynoothuniversity.ie/sites/default/files/assets/document//Maynooth%20University%20%20Research%20Ethics%20Policy%20%28Updated%20March%202020%29.pdf. [URL Memento: Wayback Machine]

Maynooth University. (2021). Maynooth University Research Integrity Policy. Maynooth University. Retrieved 2021-10-05, from https://www.maynoothuniversity.ie/sites/default/files/assets/document//MU%20Research%20Integrity%20%20Policy%20V4.0%2026%2004%20%2021_approved%20by%20Research%20Committee.pdf. [URL Memento: Wayback Machine]

Meyer, E. T., Thomas, A., & Schroeder, R. (2011). *Web Archives: The Future(s)*. Oxford Internet Institute; International Internet Preservation Consortium. Retrieved 2021-09-10, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1830025. DOI: 10.2139/ssrn.1830025

Michel, A., Pranger, J., Geeraert, F., Lieber, S., Mechant, P., Vlassenroot, E., Chambers, S., Birkholz, J., & Messens, F. (2021). *WP1 report: An international review of Social Media Archiving initiatives* [BESOCIAL – Report WP1]. Retrieved 2022-07-20, from https://orfeo.belnet.be/handle/internal/7741. [URL Memento: Wayback Machine]

Microsoft. (n.d.). Microsoft Excel Spreadsheet Software | Microsoft 365 [Web page]. Microsoft. https://www.microsoft.com/en-ie/microsoft-365/excel. [URL Memento: archive.today]

Milligan, I. (2015, July 14). Web Archive Legal Deposit: A Double-Edged Sword. Ian Milligan: Digital History, Web Archives, and Contemporary History. Retrieved 2022-01-24, from https://ianmilli.wordpress.com/2015/07/14/web-archive-legal-deposit-a-double-edged-sword/. [URL Memento: Wayback Machine]

Milligan, I. (2019). *History in the Age of Abundance? How the Web is Transforming Historical Research* (Paperback). Canada: McGill-Queen's University Press.

Moiraghi, E. (2018). Le projet Corpus et ses publics potentiels. [Research Report] Une étude prospective sur les besoins et les attentes des futurs usagers. Bibliothèque nationale de France. Retrieved 2021-04-12, from https://hal-bnf.archives-ouvertes.fr/hal-01739730. [URL Memento: Wayback Machine]

Mourão, A., & Gomes, D. (2021). *The Anatomy of a Web Archive Image Search Engine—Technical Report* [Technical Report]. FCT: Arquivo.pt. Retrieved 2022-01-08, from https://sobre.arquivo.pt/wp-content/uploads/The_Anatomy_of_a_Web_Archive_Image_Search_Engine_tech_report.pdf. [URL Memento: Wayback Machine]

Murray, K. R., & Hsieh, I., K. (2008). Archiving Web-published materials: A needs assessment of librarians, researchers, and content providers. *Government Information Quarterly*, 25(1), 66–89. DOI: 10.1016/j.giq.2007.04.005

National Digital Stewardship Alliance. (2022, April 14). About the NDSA [Web page]. *N*ational Digital Stewardship Alliance - Digital Library Federation. Retrieved 2022-04-16, from http://ndsa.org//about/. [URL Memento: Wayback Machine]

National Digital Stewardship Alliance. (n.d.). National Digital Stewardship Alliance [Web page]. National Digital Stewardship Alliance - Digital Library Federation. Retrieved 2021-11-24, from http://ndsa.org/. [URL Memento: Wayback Machine]

National Digital Stewardship Alliance Content Working Group. (2012). *Web Archiving Survey Report* [NDSA Report]. USA: National Digital Stewardship Alliance (NDSA). Retrieved 2022-03-19, from https://osf.io/na24q/. [URL Memento: archive.today]

National LIbrary of Australia. (n.d.). PANDORA Web Archive [Website]. PANDORA Web Archive, National LIbrary of Australia. Retrieved 2022-01-24, from http://pandora.nla.gov.au. [URL Memento: Wayback Machine]

NetLab (n.d.). Tools and Tutorials | NetLab [Web page]. NetLab. Retrieved 2021-04-26, from https://www.netlab.dk/services/tools-and-tutorials/. [URL Memento: Wayback Machine]

Newing, C., & Clegg, P. (2021, February 9). Making the UK Government Social Media Archive even better [Blog post]. IIPC Blog. Retrieved 2021-11-04, from https://netpreserveblog.wordpress.com/2021/02/09/making-the-uk-government-social-media-archive-even-better/. [URL Memento: Wayback Machine]

Nielsen, J. (2016). *Using Web Archives in Research—An Introduction* (Online/pdf). Aarhus, Denmark: NetLab. Retrieved 2022-06-22, from https://dighumlab.org/wp-content/uploads/2017/06/Nielsen_Using_Web_Archives_in_Research.pdf. [URL Memento: archive.today]

Niu, J. (2012). An Overview of Web Archiving. *D-Lib Magazine*, 18(3/4). DOI: 10.1045/march2012-niu1 [URL Memento: Wayback Machine]

Ogden, J. (2021). "Everything on the internet can be saved": Archive Team, Tumblr and the cultural significance of web archiving. *Internet Histories*, 0(0), 1–20. DOI: 10.1080/24701475.2021.1985835

Old Dominion University, Web Science and Digital Libraries Research Group at Old Dominion University. (n.d.). Web Science and Digital Libraries Research Group (ODU WS-DL) [Blog site]. Web Science and Digital Libraries Research Group. Retrieved 2021-07-29, from https://ws-dl.blogspot.com/. [URL Memento: Wayback Machine]

Pennock, M. (2013). *Web-Archiving* [Report]. DPC Technology Watch Report 13. UK: Digital Preservation Coalition in association with Charles Beagrie Ltd. Retrieved 2019-02-08, from https://www.dpconline.org/docs/technology-watch-reports/865-dpctw13-01-pdf/file. DOI: 10.7207/twr13-01 [URL Memento: Wayback Machine]

Post, C. (2017). Building a Living, Breathing Archive: A Review of Appraisal Theories and Approaches for Web Archives. *Preservation, Digital Technology & Culture*, 46(2), 69–77. DOI: 10.1515/pdtc-2016-0031

QualCoder. (n.d.). QualCoder [Website]. QualCoder. Retrieved 2022-05-20, from https://qualcoder.wordpress.com. [URL Memento: Wayback Machine]

Quint, B. (1998, October 19). A 'Gift of the Web' for the Library of Congress from Alexa Internet. *Information Today, Newsbreaks*, [online]. Retrieved 2022-01-23, from http://newsbreaks.infotoday.com/NewsBreaks/A-Gift-of-the-Web-for-the-Library-of-Congress-from-Alexa-Internet-17893.asp. [URL Memento: Wayback Machine]

Ras, M., & van Bussel, S. (2007). *Web Archiving User Survey* [Technical Report]. National Library of the Netherlands (Koninklijke Bibliotheek). Retrieved from http://web.archive.org/web/20220120040514/https://www.kb.nl/sites/default/files/docs/kb_usersurvey_webarchive_en.pdf. [Web archive: Wayback Machine; source URL: https://www.kb.nl/sites/default/files/docs/kb_usersurvey_webarchive_en.pdf; Timestamp: 2022-01-20 04:05:14]

Recite Me. (n.d.). Recite Me: Choosing an Accessible Font [Web page/pdf]. Recite Me. Retrieved 2021-12-16, from

https://reciteme.com/uploads/articles/accessible_fonts_guide.pdf. [URL Memento: Wayback Machine]

RESAW. (n.d.). About RESAW | RESAW [Web page]. Research Infrastructure for the Study of Archived Web Materials. Retrieved 2021-03-10, from http://resaw.eu/about/. [URL Memento: Wayback Machine]

Riley, H., & Crookston, M. (2015). *Awareness and Use of the New Zealand Web Archive: A survey of New Zealand academics* [Report]. New Zealand: University of Wellington; National Library of New Zealand. Retrieved 2018-01-06, from https://natlib.govt.nz/files/webarchive/nzwebarchive-awarenessanduse.pdf. [URL Memento: Wayback Machine]

Rosenthal, D. (2015, November 19). You get what you get and you don't get upset [Blog post]. DSHR's Blog. Retrieved 2021-05-29, from https://blog.dshr.org/2015/11/you-get-what-you-get-and-you-dont-get.html. [URL Memento: Wayback Machine]

Rosnay, M. D. de, Georges, F., Crosnier, H. L., Merzeau, L., Musiani, F., Paloque-Berges, C., Schafer, V., & Thierry, B. (Eds.) (n.d.). Web90 – Heritage, Memories and History of the Web in the 1990s | Home [Blog site]. Web90 project. Retrieved 2022-07-09, from https://web90.hypotheses.org/. [URL Memento: Wayback Machine]

Ruest, N., Lin, J., Milligan, I., & Fritz, S. (2020). The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. Association for Computing Machinery, New York, NY, USA, 157–166. DOI: 10.1145/3383583.3398513

Ryan, M., Keating, D., & Finegan, J. (2022). Managing and accessing web archives: Irish practitioners' perspectives. *AI & SOCIETY*. DOI: 10.1007/s00146-021-01364-0 [URL Memento: Wayback Machine]

Samar, T., Traub, M. C., van Ossenbruggen, J., Hardman, L., & de Vries, A. P. (2017). Quantifying retrieval bias in Web archive search. *International Journal on Digital Libraries*, 19(1), 57–75. DOI: 10.1007/s00799-017-0215-9 [URL Memento: Wayback Machine]

Schafer, V. & Winters, J. (2021). The values of web archives. *International Journal of Digital Humanities*, 2(1), 129–144. DOI: 10.1007/s42803-021-00037-0 [URL Memento: Wayback Machine]

Schneider, S. M., Foot, K. A., & Wouters, P. (2009). Web archiving as e- research. In N. W. Jankowski (Ed.), *E-Research: Transformation in Scholarly Practice* (pp. 205–221). New York, London: Routledge.

Schroeder, R. & Brügger, N. (2017). Introduction: The Web as History. In Niels Brügger & R. Schroeder (Eds.), *The Web as History: Using Web Archives to Understand the Past and the Present* (Online/pdf, pp. 1–22). London: UCL Press. DOI: 10.14324/111.9781911307563 [URL Memento: Wayback Machine]

Shankland, S. (2009, April 23). Now closing: GeoCities, a relic of Web's early days [News article]. CNET. Retrieved 2022-01-24, from https://www.cnet.com/tech/services-

and-software/now-closing-geocities-a-relic-of-webs-early-days/. [URL Memento: Wayback Machine]

Society of American Archivists. (2005). Provenance. In *Dictionary of Archives Terminology* (Online/web). Society of American Archivists (SAA). Retrieved 2021-07-30, from https://dictionary.archivists.org/entry/provenance.html. [URL Memento: Wayback Machine]

Spaniol, M., Denev, D., Mazeika, A., Weikum, G., & Senellart, P. (2009). Data Quality in Web Archiving. *Proceedings of the 3rd Workshop on Information Credibility on the Web*, 19–26. DOI: 10.1145/1526993.1526999

Spinellis, D. (2003). The Decay and Failures of Web References. *Communications of the ACM*, 46(1), 71–77. DOI: 10.1145/602421.602422

Steber, C. (2016). Online Surveys: Data Collection Advantages & Disadvantages [Blog post]. *Communications for Research*. Retrieved from https://web.archive.org/web/20201003184529/https://www.cfrinc.net/cfrblog/online-surveys-advantages-disadvantages. [Web archive: Wayback Machine; source URL: https://www.cfrinc.net/cfrblog/online-surveys-advantages-disadvantages; Timestamp: 2020-10-03 18:45:29]

Summers, E. (2020). Appraisal Talk in Web Archives. *Archivaria*, 89(1), 70–102. Project Muse: http://muse.jhu.edu/article/755769. Project Muse ID: 755769

Summers, E., & Punzalan, R. (2017). Bots, Seeds and People: Web Archives as Infrastructure. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 821–834. DOI: 10.1145/2998181.2998345

Taguette. (n.d.). Taguette, the free and open-source qualitative data analysis tool [Website]. *Taguette*. Retrieved from https://www.taguette.org/. [URL Memento: Wayback Machine]

Taylor, N. (2017). Understanding Legal Use Cases for Web Archives [Conference presentation]. *International Internet Preservation Coalition General Assembly and Web Archiving Conference, London, June 2016*. https://nullhandle.org/pdf/2017-06-16_understanding_legal_use_cases_for_web_archives.pdf. [URL Memento: Wayback Machine]

Teleport. (n.d.). Teleport [Website]. Teleport. Retrieved 2021-12-29, from https://goteleport.com. [URL Memento: Wayback Machine]

The Archives Unleashed Project. (n.d.). The Archives Unleashed Project [Website]. The Archives Unleashed Project. Retrieved 2021-11-23, from https://archivesunleashed.org. [URL Memento: Wayback Machine]

The Archives Unleashed Project. (n.d.). Archives Unleashed Cohorts (2022-2023) [Web page]. The Archives Unleashed Project. Retrieved 2022-02-17, from https://archivesunleashed.org/cohorts2022-2023/. [URL Memento: Wayback Machine]

The Bodleian Libraries. (n.d.). Legal deposit [Web page]. The Bodleian Libraries, University of Oxford. Retrieved 2022-02-03, from https://www.bodleian.ox.ac.uk/collections-and-resources/legal-deposit. [URL Memento: Wayback Machine]

Thomas, A., Meyer, E. T., Dougherty, M., van den Heuvel, C., Madsen, C. M., & Wyatt, S. (2010). *Researcher Engagement with Web Archives: Challenges and Opportunities for Investment*. Joint Information Systems Committee Report, August 2010. UK: JISC. Retrieved 2020-04-14, from https://papers.ssrn.com/abstract=1715000. [URL Memento: Wayback Machine]

Truman, G. (2016). *Web Archiving Environmental Scan* [Report]. Harvard Library Report, 2016. USA: Harvard Library. Retrieved 2021-02-24, from https://dash.harvard.edu/handle/1/25658314. [URL Memento: Wayback Machine]

Truter, V. (2021). Research Data Management and Sharing Practices of Researchers in Web Archive Studies [Conference presentation & abstract]. *Engaging with Web Archives 4 Digital Humanities (#EWA4DH), Maynooth University Arts and Humanities Institute, Co. Kildare, Ireland, 21 September 2021*. Retrieved 2022-05-07, from https://ewaconference.com/ewa4dh-2021/ewa4dh-programme/. [URL Memento: Wayback Machine]

UK Web Archive. (2018, February 1). A New Playback Tool for the UK Web Archive [Blog post]. UK Web Archive Blog. Retrieved 2021-05-12, from https://blogs.bl.uk/webarchive/2018/02/index.html. [URL Memento: UK Web Archive]

UK Parliament. (2013) The Legal Deposit Libraries (Non-Print Works) Regulations 2013, UK Statutory Instruments 2013 No. 777. Retrieved 2022-03-08, from https://www.legislation.gov.uk/uksi/2013/777/contents/made. [URL Memento: Wayback Machine]

Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., & Mechant, P. (2019). Web archives as a data resource for digital scholars. *International Journal of Digital Humanities*, 1(1), 85–111. DOI: 10.1007/s42803-019-00007-7

Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J., & Mechant, P. (2021). Web-archiving and social media: An exploratory analysis. *International Journal of Digital Humanities*, 2(1), 107–128. DOI: 10.1007/s42803-021-00036-1 [URL Memento: Wayback Machine]

WARCnet. (n.d.). Aarhus Autumn 2021 | WARCnet [Web page]. Aarhus University. Retrieved 2022-05-07, from https://cc.au.dk/en/warcnet/meetings/aarhus-autumn-2021. [URL Memento: Wayback Machine]

WARCnet. (n.d.). About WARCnet | WARCnet [Web page]. Aarhus University. Retrieved 2021-04-23, from https://cc.au.dk/en/warcnet/about/. [URL Memento: Wayback Machine]

Webster, P. (2017). Users, technologies, organisations: Towards a cultural history of world web archiving. In N. Brügger (Ed.), *Web 25: Histories from the first 25 Years of the World Wide Web* (pp. 175–190). New York: Peter Lang.

Webster, P. (2020). *How Researchers Use the Archived Web*. DPC Technology Watch Guidance Note. UK: Digital Preservation Coalition. Retrieved 2021-12-07, from https://www.dpconline.org/docs/technology-watch-reports/2263-twgn-20-01-how-researchers-use-the-archived-web-webster/file. DOI 10.7207/twgn20-01 [URL Memento: Wayback Machine]

Wikipedia. (2002). Information science (also known as information studies). *Wikipedia*. Retrieved 2022-02-02, from https://en.wikipedia.org/w/index.php?title=Information_science&oldid=1040622798. [URL Memento: archive.today]

Wikipedia. (2011+). List of Web archiving initiatives. *Wikipedia*. Retrieved 2021-09-18, from https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives. [URL Memento: Wayback Machine]

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

Winters, J. (2020a). Giving with one Click, Taking with the Other: Electronic Legal Deposit, Web Archives and Researcher Access. In M. Terras & P. Gooding (Eds.), *Electronic Legal Deposit: Shaping the Library Collections of the Future* (pp. 159–178). London: Facet Publishing. DOI: 10.29085/9781783303786.010

Winters, J. (2020b). Web archives as sites of collaboration [Conference keynote presentation (video)]. *Engaging with Web Archives: 'Opportunities, Challenges and Potentialities', (#EWAVirtual), Maynooth University Arts and Humanities Institute, Co. Kildare, Ireland, [online], 21-22 September 2022*. Retrieved 2021-09-18, from https://www.youtube.com/watch?v=c5JYCfnLJ-c

Winters, J., & Prescott, A. (2019). Negotiating the born-digital: A problem of search. *Archives and Manuscripts*, 47(3), 391–403. DOI: 10.1080/01576895.2019.1640753

Wright, R. (1997, May 19). Tim Berners-Lee: The Man Who Invented the Internet [Magazine online]. *Time*, 149(20), p. 1/7. Retrieved 2017-09-21, from http://content.time.com/time/subscriber/article/0,33009,986354,00.html. [URL Memento: Wayback Machine]

Xie, Z., Klein, M., & Fox, E. A. (2020). Web Archiving and Digital Libraries. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 583–584. DOI: 10.1145/3383583.3398509

# Providers & Services


Archive-It (n.d.). Archive-It—Web Archiving Services for Libraries and Archives [Website]. Archive-It (Internet Archive). Retrieved 2021-03-16, from https://archive-it.org. [URL Memento: Wayback Machine]

archive.today. (n.d.). archive.today: Webpage capture [Website]. archive.today. Retrieved 2021-03- from https://archive.ph. [URL Memento: Wayback Machine]

Arquivo.pt. (n.d.). Arquivo.pt—Pesquise páginas do passado! [Website]. Arquivo.pt. Retrieved 2021-03-23, from https://arquivo.pt. [URL Memento: Arquivo.pt]

Austrian National Library. (n.d.). Webarchiv Österreich [Website]. Webarchiv Österreich. Retrieved 2021-03-10, from https://webarchiv.onb.ac.at. [URL Memento: Wayback Machine]

Biblioteca Nacional de España. (n.d.). Archivo de la Web Española [Web page]. Biblioteca Nacional de España. Retrieved 2021-09-09, from http://www.bne.es/es/Colecciones/ArchivoWeb. [URL Memento: Wayback Machine]

Bibliothèque nationale de France. (n.d.). BnF Archives de l'internet [Web page]. BnF; Bibliothèque nationale de France. Retrieved 2021-01-28, from https://www.bnf.fr/fr/archives-de-linternet. [URL Memento: Wayback Machine]

Bibliothèque nationale du Luxembourg. (n.d.). Luxembourg Web Archive – WEBARCHIVE.LU [Web page]. Bibliothèque Nationale Du Luxembourg. Retrieved 2021-03-11, from https://www.webarchive.lu. [URL Memento: Wayback Machine]

Common Crawl. (n.d.). Common Crawl [Website]. Common Crawl. Retrieved 2021-03-12, from https://commoncrawl.org. [URL Memento: Wayback Machine]

Det Kgl. Bibliotek. (n.d.). Netarkivet [Web page]. Det Kgl. Bibliotek. Retrieved 2021-03-09, from https://www.kb.dk/en/find-materials/collections/netarkivet. [URL Memento: Wayback Machine]

Institut national de l'audiovisuel. (n.d.). Institut national de l'audiovisuel (INA) [Website]. Institut national de l'audiovisuel. Retrieved 2021-10-08, from https://www.ina.fr/. [URL Memento: Wayback Machine]

International Internet Preservation Consortium. (n.d.). International Internet Preservation Consortium [Website]. International Internet Preservation Consortium. Retrieved 2021-04-22, from https://netpreserve.org. [URL Memento: Wayback Machine]

Internet Archive. (n.d.). Wayback Machine (Internet Archive) [Web page]. Internet Archive. Retrieved 2021-02-02, from https://archive.org/web/. [URL Memento: Wayback Machine]

Koninklijke Bibliotheek. (n.d.). Web archiving [Web page]. Koninklijke Bibliotheek. Retrieved 2022-06-28, from https://www.kb.nl/en/about-us/expertise/web-archiving. [URL Memento: Wayback Machine]

Library and Archives Canada. (n.d.). Government of Canada Web Archive [Web page]. Library and Archives Canada. Retrieved 2021-03-08, from https://www.bac-lac.gc.ca/eng/discover/archives-web-government/Pages/web-archives.aspx. [URL Memento: Wayback Machine]

Library of Congress. (n.d.). Library of Congress Web Archives [Web page]. Library of Congress. Retrieved 2021-03-26, from https://www.loc.gov/web-archives/collections. [URL Memento: Wayback Machine]

Memento Project. (n.d.). Memento Time Travel [Website]. Memento Time Travel. Retrieved 2021-03-22, from http://timetravel.mementoweb.org. [URL Memento: Wayback Machine]
National and University Library in Zagreb. (n.d.). Croatian Web Archive (HAW) [Website]. Croatian Web Archive. Retrieved 2021-10-08, from https://haw.nsk.hr/en. [URL Memento: Wayback Machine]
National Library of Ireland. (n.d.). National Library of Ireland Web Archive | Archive-IT [Web page]. Archive-It. Retrieved 2021-03-01, from https://archive-it.org/home/nli. [URL Memento: Wayback Machine]
National Records of Scotland. (n.d.). National Records of Scotland Web Archive [Website]. National Records of Scotland Web Archive. Retrieved 2021-11-03, from https://webarchive.nrscotland.gov.uk/#!/. [URL Memento: Wayback Machine]
OSZK Webarchívum. (n.d.). OSZK Webarchívum. OSZK Webarchívum (National Széchényi Library). Retrieved 2021-05-19, from https://webarchivum.oszk.hu/en/for-users/short-description/. [URL Memento: Wayback Machine]
Public Records Office of Northern Ireland (PRONI). (n.d). PRONI Web Archive [Web page]. NI Direct Government Services. Retrieved 2021-02-22, from https://www.nidirect.gov.uk/services/search-proni-web-archive. [URL Memento: Wayback Machine]
Rhizome. (n.d.). Conifer (previously Webrecorder) [Website]. Rhizome. Retrieved 2021-05-19, from https://conifer.rhizome.org. [URL Memento: Wayback Machine]
Sherratt, T. & Andrew Jackson. (2021). GLAM Workbench—Web Archives [Wiki]. GLAM Workbench/GitHub.Io. Retrieved 2022-04-29, from https://glam-workbench.github.io/web-archives. [URL Memento: Wayback Machine]
The National Archives. (n.d.). UK Government Web Archive [Web page]. The National Archives (UK). Retrieved 2021-03-17, from http://www.nationalarchives.gov.uk/webarchive. [URL Memento: Wayback Machine]
UK Parliamentary Archives. (n.d.). UK Parliament Web Archive [Website]. UK Parliament Web Archive. Retrieved 2021-02-02, from http://webarchive.parliament.uk/. [URL Memento: Wayback Machine]
UK Web Archive. (2013). JISC UK Web Domain Dataset (1996-2010) [Web page]. UK Web Archive. Retrieved 2021-11-02, from http://data.webarchive.org.uk/opendata/ukwa.ds.2/. [URL Memento: Wayback Machine] DOI: 10.5259/UKWA.DS.2/1
UK Web Archive. (n.d.a). SHINE [Web page]. UK Web Archive. Retrieved 2021-11-14, from https://www.webarchive.org.uk/shine. [URL Memento: Wayback Machine]; QID: None
UK Web Archive. (n.d.b). UK Web Archive [Website]. UK Web Archive. Retrieved 2021-03-18, from https://www.webarchive.org.uk/ukwa. [URL Memento: Wayback Machine]
Zone-H. (n.d.). Zone-H: Unrestricted information [Website]. Zone-H Unrestricted Information. Retrieved 2021-10-14, from http://www.zone-h.org/?hz=1. [URL Memento: Wayback Machine]

# Software, Tools & Methods


Amazon Web Services. (n.d.). Amazon Athena [Web page]. Amazon Web Services, Inc. Retrieved 2021-12-16, from https://aws.amazon.com/athena/. [URL Memento: Wayback Machine]

Amazon Web Services. (n.d.b). Amazon Web Services (AWS) [Website]. Amazon Web Services, Inc. Retrieved 2021-11-30, from https://aws.amazon.com/. [URL Memento: Wayback Machine]

Apache Software Foundation. (2011). Apache Lucene [Website]. Apache Lucene. Retrieved 2021-10-29, from https://lucene.apache.org/index.html. [URL Memento: Wayback Machine]

Apache Software Foundation. (n.d.). Apache Parquet [Website]. Apache Parquet. Retrieved 2021-12-28, from https://parquet.apache.org/. [URL Memento: Wayback Machine]

Apache Software Foundation. (n.d.). Solr [Website]. Apache Solr. Retrieved 2021-10-29, from https://solr.apache.org/index.html. [URL Memento: Wayback Machine]

ATLAS.ti. (n.d.). ATLAS.ti: The Qualitative Data Analysis & Research Software [Website]. ATLAS.Ti. Retrieved 2022-01-22, from https://atlasti.com/. [URL Memento: Wayback Machine]

Atlassian. (n.d.). Confluence Data Center and Server support. Atlassian Support. Retrieved 2021-12-20, from https://support.atlassian.com/confluence-server/. [URL Memento: Wayback Machine]

Bibliotheca Alexandrina. (2020+). Link-indexer | LinkGate [Software]. Bibliotheca Alexandrina. Retrieved 2022-04-30, from https://github.com/arcalex/link-indexer. [URL Memento: Wayback Machine]

Bibliotheca Alexandrina. (2020+). Link-serv | LinkGate [GitHub]. Bibliotheca Alexandrina. Retrieved 2022-02-24, from https://github.com/arcalex/link-serv. [URL Memento: Wayback Machine]

Bibliotheca Alexandrina. (2020+). Link-viz | LinkGate [Software]. Retrieved 2022-02-24, from https://github.com/arcalex/link-viz. [URL Memento: Wayback Machine]

Bibliotheca Alexandrina. (2020+). LinkGate [GitHub]. Bibliotheca Alexandrina. Retrieved 2022-04-30, from https://github.com/arcalex/linkgate. [URL Memento: archive.today]

BibTeX. (n.d.). BibTeX [Website]. BibTeX. Retrieved 2021-10-05, from http://www.bibtex.org. [URL Memento: Wayback Machine]

BitCurator NLP. (n.d.). BitCurator [Website]. BitCurator. Retrieved 2021-12-24, from https://bitcurator.net/. [URL Memento: Wayback Machine]

Blue Squirrel. (n.d.). [Grab-a-Site] offline cd browser Grab-A-Site 5.0 software for windows [Web page]. Blue Squirrel. Retrieved 2021-04-12, from https://www.bluesquirrel.com/products/grabasite/. [URL Memento: Wayback Machine]

Documenting the Now. (2013+). Twarc [GitHub]. Documenting the Now (DocNow/twarc). Retrieved 2021-07-06, from https://github.com/DocNow/twarc. [URL Memento: Wayback Machine]

DSpace. (n.d.). DSpace | Home [Website]. DSpace. https://dspace.lyrasis.org/. [URL Memento: Wayback Machine]

Egense, T. (2017+). SolrWayback [GitHub]. NetarchiveSuite (netarchivesuite/solrwayback). Retrieved 2021-07-06, from https://github.com/netarchivesuite/solrwayback. [URL Memento: Wayback Machine]
Elasticsearch. (n.d.). Elastic Stack: Elasticsearch, Kibana, Beats & Logstash [Web page]. Elasticsearch. Retrieved 2021-11-23, from https://www.elastic.co/elastic-stack. [URL Memento: Wayback Machine]
Elasticsearch (n.d.). Elasticsearch: The Official Distributed Search & Analytics Engine [Web page]. Elasticsearch. Retrieved 2021-12-07, from https://www.elastic.co/elasticsearch. [URL Memento: Wayback Machine]
Elasticsearch. (n.d.). Kibana: Explore, Visualize, Discover Data | Elastic [Web page]. Elasticsearch. Retrieved 2021-12-22, from https://www.elastic.co/kibana/. [URL Memento: Wayback Machine]
Gephi. (2011+). Gephi [GitHub]. Gephi (gephi/gephi). Retrieved 2021-07-13, from https://github.com/gephi/gephi. [URL Memento: Wayback Machine]
Gephi. (n.d.). Gephi—The Open Graph Viz Platform [Website]. Gephi. Retrieved 2021-08-12, from https://gephi.org/. [URL Memento: Wayback Machine]
GNU Project & Free Software Foundation. (n.d.). Wget - GNU Project [Web page]. GNU.org. Retrieved 2021-10-10, from https://www.gnu.org/software/wget/. [URL Memento: Wayback Machine]
Graf, A. (n.d.). Instaloader—Download Instagram Photos and Metadata [Wiki]. Instaloader/GitHub.io. Retrieved 2021-06-07, from https://instaloader.github.io/. [URL Memento: archive.today]
GW Libraries and Academic Innovation. (2015+). Social Feed Manager/sfm-ui [GitHub]. GW Libraries and Academic Innovation (gwu-libraries/sfm-ui). Retrieved 2021-04-14, from https://github.com/gwu-libraries/sfm-ui. [URL Memento: Wayback Machine]
GW Libraries and Academic Innovation. (n.d.). Social Feed Manager [Web page]. Social Feed Manager/GitHub.io. Retrieved 2021-04-15, from https://gwu-libraries.github.io/sfm-ui/. [URL Memento: Wayback Machine]
HeidiSQL. (n.d.). HeidiSQL - MariaDB, MySQL, MSSQL, PostgreSQL and SQLite made easy [Website]. HeidiSQL.com. Retrieved 2021-12-14, from https://www.heidisql.com/. [URL Memento: Wayback Machine]
International Federation of Library Associations and Institutions. (n.d.). International Standard Bibliographic Description (ISBD)—IFLA. International Federation of Library Associations and Institutions. Retrieved 2022-05-06, from https://www.ifla.org/references/best-practice-for-national-bibliographic-agencies-in-a-digital-age/resource-description-and-standards/bibliographic-control/international-standard-bibliographic-description-isbd/. [URL Memento: Wayback Machine]
International Internet Preservation Consortium (2017+). Awesome Web Archiving [GitHub]. International Internet Preservation Consortium (iipc/awesome-web-archiving). Retrieved 2021-02-23, from https://github.com/iipc/awesome-web-archiving. [URL Memento: Wayback Machine]
Internet Archive. (2011+). Heritrix [GitHub]. Internet Archive (internetarchive/heritrix3). Retrieved 2021-11-23, from https://github.com/internetarchive/heritrix3. [URL Memento: Wayback Machine]

Internet Archive. (2013+). Wayback [GitHub]. Internet Archive (internetarchive/wayback). Retrieved 2021-11-10, from https://github.com/internetarchive/wayback. [URL Memento: Wayback Machine]
Internet Archive. (2014+). Umbra [GitHub]. Internet Archive (internetarchive/umbra). Retrieved 2021-03-11, from https://github.com/internetarchive/umbra. [URL Memento: Wayback Machine]
Internet Archive. (2015+). Brozzler [GitHub]. Internet Archive (internetarchive/brozzler). Retrieved 2021-03-13, from https://github.com/internetarchive/brozzler. [URL Memento: Wayback Machine]
Internet Archive. (n.d.b). Save Page Now—Wayback Machine | Internet Archive [Web page]. Internet Archive. Retrieved 2021-02-02, from https://archive.org/web/. [URL Memento: Wayback Machine]
ISO - International Organization for Standardization. (n.d.). ISO—International Organization for Standardization [Website]. ISO - International Organization for Standardization. Retrieved 2021-09-07, from https://www.iso.org/home.html. [URL Memento: Wayback Machine]
JISC Online Surveys. (n.d.). JISC Online Surveys [Website]. JISC Online Surveys. Retrieved 2021-04-17, from https://www.onlinesurveys.ac.uk/. [URL Memento: Wayback Machine]
Jupyter Team. (2015). The Jupyter Notebook—Jupyter Notebook 6.4.12 documentation [Wiki]. Jupyter Notebook/ReadTheDocs.Io. Retrieved 2021-10-07, from https://jupyter-notebook.readthedocs.io/en/stable/. [URL Memento: Wayback Machine]
Kelly, M. (2013+). Web Archiving Integration Layer (WAIL) [GitHub]. Mat Kelly (machawk1/wail). Retrieved 2021-11-23, from https://github.com/machawk1/wail. [URL Memento: Wayback Machine]
Kelly, M. (n.d.). WAIL—Web Archiving Integration Layer [Web page]. WAIL/GitHub.io. Retrieved 2021-10-06, from http://machawk1.github.io/wail/. [URL Memento: Wayback Machine]
Kreymer, I. (2013+). Pywb (Webrecorder pywb 2.6) [GitHub]. Webrecorder (webrecorder/pywb). Retrieved 2021-07-06, from https://github.com/webrecorder/pywb. [URL Memento: Wayback Machine]
Kreymer, I. (2019+). Browsertrix [GitHub]. Webrecorder (webrecorder/browsertrix-cloud). Retrieved 2021-03-25, from https://github.com/webrecorder/browsertrix. [URL Memento: Wayback Machine]
Kreymer, I. (2020+). OldWeb.today [GitHub]. OldWeb.today (oldweb-today). Retrieved 2021-03-26, from https://github.com/oldweb-today/oldweb-today. [URL Memento: Wayback Machine]
Kreymer, I. (n.d.). Pywb | Webrecorder pywb documentation! [Website]. Pywb/Readthedocs.io. Retrieved 2021-10-29, from https://pywb.readthedocs.io/en/latest/. [URL Memento: Wayback Machine]
Mahanty, A. (2020). Waybackpy [GitHub]. Akash Mahanty (akamhy/waybackpy). Retrieved 2021-10-07, from https://github.com/akamhy/waybackpy. [URL Memento: Wayback Machine]
Mahanty, A. (n.d.). Waybackpy [Web page]. waybackpy/GitHub.io. Retrieved 2021-01-05, from https://akamhy.github.io/waybackpy/. [URL Memento: Wayback Machine]

MathWorks. (n.d.). MATLAB | MathWorks [Web page]. MathWorks. Retrieved , 2021-06-21 from https://uk.mathworks.com/products/matlab.html. [URL Memento: Wayback Machine]
MAXQDA. (n.d.). MAXQDA | All-In-One Qualitative & Mixed Methods Data Analysis Tool [Website]. MAXQDA. Retrieved 2021-08-18, from https://www.maxqda.com. [URL Memento: Wayback Machine]
MediaArea. (n.d.). MediaArea [Website]. MediaArea. Retrieved 2021-10-20, from https://mediaarea.net. [URL Memento: Wayback Machine]
Memento Project. (n.d.). Memento Time Travel [Website]. Memento Time Travel. Retrieved 2021-03-22, from http://timetravel.mementoweb.org/. [URL Memento: Wayback Machine]
Microsoft. (n.d.). Microsoft Excel Spreadsheet Software | Microsoft 365 [Web page]. Microsoft. Retrieved 2021-05-24, from https://www.microsoft.com/en-ie/microsoft-365/excel. [URL Memento: archive.today]
Microsoft. (n.d.). Power Pivot—Overview and Learning | Microsoft [Web page]. Microsoft. Retrieved 2021-02-12, from https://support.microsoft.com/en-us/office/power-pivot-overview-and-learning-f9001958-7901-4caa-ad80-028a6d2432ed. [URL Memento: archive.today]
Microsoft. (n.d.). PowerShell Documentation—PowerShell [Web page]. Microsoft. Retrieved 2021-06-15, from https://docs.microsoft.com/en-us/powershell/. [URL Memento: Wayback Machine]
MirrorWeb. (n.d.). SEC 17a-4 (Electrolyte) [Web Page]. MirrorWeb. Retrieved 2022-08-10, from https://www.mirrorweb.com/solutions/sec-17a-4 .[URL Memento: Wayback Machine]
National Library of the Netherlands & National Library of New Zealand. (n.d.). Web Curator Tool [Website]. Web Curator Tool. Retrieved 2021-10-17, from https://webcuratortool.org. [URL Memento: Wayback Machine]
Netarkivet.dk. (2014+). NetarchiveSuite | Introduction [GitHub]. NetarchiveSuite (netarchivesuite). Retrieved 2022-04-30, from https://github.com/netarchivesuite/netarchivesuite. [URL Memento: Wayback Machine] original-date: 2014-05-15T12:23:27Z
Netarkivet.dk, Sørensen, M. S., & Have, U. K. (n.d.). NetarchiveSuite [Web page]. NetarchiveSuite/SBForge.org. Retrieved 2021-11-26, from https://sbforge.org/display/NAS. [URL Memento: Wayback Machine]
Nutchwax. (2005-2009). Nutchwax [Web page]. Sourceforge.net. Retrieved 2021-07-11, from http://archive-access.sourceforge.net/projects/nutchwax. [URL Memento: Wayback Machine]
Old Dominion University Web Science and Digital Libraries Research Group. (2017+). Archive Now [GitHub]. Old Dominion University Web Science and Digital Libraries Research Group (oduwsdl/archivenow). Retrieved 2021-01-13, from https://github.com/oduwsdl/archivenow. [URL Memento: Wayback Machine]
Old Dominion University Web Science and Digital Libraries Research Group (n.d.). Dark and Stormy Archives: Storytelling with web archive collections [Web page]. Dark and Stormy Archives/GitHub.io. Retrieved 2021-10-01, from https://oduwsdl.github.io/dsa/. [URL Memento: Wayback Machine]
OpenRefine. (n.d.). OpenRefine [Website]. OpenRefine. Retrieved 2021-08-29, from https://openrefine.org/. [URL Memento: Wayback Machine]

OpenWayback Development. (2012+). OpenWayback [GitHub]. OpenWayback Development/International Internet Preservation Consortium (iipc/openwayback). Retrieved 2022-05-31, from https://github.com/iipc/openwayback. [URL Memento: Wayback Machine]

OpenWayback Development. (2012+). OpenWayback | Wiki. OpenWayback Development (iipc/openwayback). Retrieved from https://github.com/iipc/openwayback/wiki. [URL Memento: Wayback Machine]

Oracle. (n.d.). MySQL [Website]. MySQL. Retrieved 2021-08-24, from https://www.mysql.com. [URL Memento: Wayback Machine]

pandas. (2010+). pandas: Powerful Python data analysis toolkit [GitHub]. pandas (pandas-dev/pandas). Retrieved 2021-12-18, from https://github.com/pandas-dev/pandas. [URL Memento: Wayback Machine]

pandas. (n.d.). pandas—Python Data Analysis Library [Website]. pandas.pydata. Retrieved 2022-04-20, from https://pandas.pydata.org. [URL Memento: Wayback Machine]

Python Software Foundation. (n.d.). Python [Website]. Python. Retrieved 2021-12-28, from https://www.python.org. [URL Memento: Wayback Machine]

QSR International. (n.d.). NVivo [Web page]. QSR International. Retrieved 2021-03-25, from https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home. [URL Memento: Wayback Machine]

Ratinaud, P. (n.d.). IRaMuTeQ [Website]. Iramuteq. Retrieved 2021-10-24, from http://www.iramuteq.org/. [URL Memento: Wayback Machine]

Rhizome. (n.d.). Conifer (previously Webrecorder). Conifer. Retrieved 2021-05-19, from https://conifer.rhizome.org. [URL Memento: Wayback Machine]

Roche, X. (2017). HTTrack Website Copier [Website]. HTTrack Website Copier. Retrieved 2021-12-30, from http://www.httrack.com. [URL Memento: Wayback Machine] versionNumber: 3.49-2 (05/20/2017)

RStudio. (n.d.). RStudio [Website]. RStudio. Retrieved 2021-10-22, from https://www.rstudio.com/. [URL Memento: Wayback Machine]

Rudis, B. (2018, September 18). Intro to the 'CDX Basic Query' Interface [Web page]. wayback/GitHub.io. Retrieved 2018-09-18, from https://hrbrmstr.github.io/wayback/articles/intro-to-cdx-basic-query.html. [URL Memento: Wayback Machine]

Sherratt, T. & Andrew Jackson. (2021). GLAM Workbench—Web Archives [Wiki]. GLAM Workbench/GitHub.io. Retrieved 2022-04-29, from https://glam-workbench.github.io/web-archives. [URL Memento: Wayback Machine]

Stéfan Sinclair & Geoffrey Rockwell. (n.d.). Voyant Tools [Website]. Voyant Tools. Retrieved 2021-10-26, from https://voyant-tools.org. [URL Memento: Wayback Machine]

Tableau. (n.d.). Tableau: We're changing the way you think about data [Website]. *Tableau.* Retrieved 2021-10-26, from https://www.tableau.com/en-gb. [URL Memento: Wayback Machine]

TastyApps. (2018, October 5). WebSnapperPro | TastyApps [Web page]. TastyApps. Retrieved 2021-09-17, from http://tastyapps.net/websnapperpro.html. [URL Memento: Wayback Machine]

TechSmith. (n.d.). Snagit | Screen capture and screen recorder [Web page]. TechSmith. Retrieved 2021-12-23, from https://www.techsmith.com/screen-capture.html. [URL Memento: Wayback Machine]

TensorFlow. (n.d.). TensorFlow [Website]. TensorFlow. Retrieved 2021-11-01, from https://www.tensorflow.org. [URL Memento: Wayback Machine]
The Archives Unleashed Project. (n.d.). The Archives Unleashed Cloud [Web page]. The Archives Unleashed Project. Retrieved 2021-05-30, from https://archivesunleashed.org/cloud. [URL Memento: Wayback Machine]
The Archives Unleashed Project. (n.d.). The Archives Unleashed Toolkit [Web page]. *The Archives Unleashed Project*. Retrieved 2021-10-21, from https://archivesunleashed.org/aut. [URL Memento: Wayback Machine]
The LaTex Project. (n.d.). LaTeX [Website]. The LaTex Project. Retrieved 2021-12-23, from https://www.latex-project.org. [URL Memento: Wayback Machine]
The National Archives. (n.d.). DROID: file format identification tool [Web page]. The National Archives. https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/. [URL Memento: Wayback Machine]
The R Foundation. (n.d.). R: The R Project for Statistical Computing | R project [Website]. The R Project. Retrieved 2021-10-12, from https://www.r-project.org. [URL Memento: Wayback Machine]
UK Web Archive. (2013+). W3Act [GitHub]. UK Web Archive. Retrieved 2021-11-23, from https://github.com/ukwa/w3act. [URL Memento: Wayback Machine]
UK Web Archive. (2013+). W3ACT [Web page]. UK Web Archive. Retrieved from https://www.webarchive.org.uk/act/login. [URL Memento: Wayback Machine]
UK Web Archive. (n.d.). SHINE | UK Web Archive [Web page]. UK Web Archive. Retrieved 2021-11-14, from https://www.webarchive.org.uk/shine. [URL Memento: Wayback Machine]
Webrecorder. (n.d.). ArchiveWeb.page [Website]. ArchiveWeb.Page. Retrieved 2022-01-05, from https://archiveweb.page. [URL Memento: Wayback Machine]
Webrecorder. (n.d.). Old Web Today [Web page]. Retrieved 2021-07-05, from https://oldweb.today/#19960101/http://geocities.com. [URL Memento: Wayback Machine]
Web Scraper. (n.d.). Web Scraper—The #1 web scraping extension [Website]. Web Scraper. Retrieved 2021-10-16, from https://webscraper.io/. [URL Memento: Wayback Machine]
Wikipedia. (2003+). Close reading. Wikipedia (Online/web). Retrieved 2021-12-26, from https://en.wikipedia.org/wiki/Close_reading. [URL Memento: Wayback Machine]
Zotero & Corporation for Digital Scholarship. (n.d.). Zotero [Website]. Zotero. Retrieved 2021-12-18, from https://www.zotero.org. [URL Memento: Wayback Machine]

# Useful Resources

Aarhus University. (n.d.). Niels Brügger—Research outputs—Aarhus University [Web page]. Aarhus University. Retrieved 2022-06-28, from https://pure.au.dk/portal/en/persons/niels-brugger(2814967c-56b1-4b7c-9599-50ff791909b7)/publications.html. [URL Memento: Wayback Machine]
Archive-It. (n.d.). Archive-It Blog [Blog site]. Archive-It. Retrieved 2021-09-01, from http://ait.blog.archive.org/. [URL Memento: Wayback Machine]

Archive-It Help Center. (n.d.). Archive-It Help Center [Web page]. Archive-It. Retrieved 2021-10-07, from https://support.archive-it.org/hc/en-us. [URL Memento: Wayback Machine]
Bibliotheca Alexandrina. (2020). LinkGate [GitHub]. Bibliotheca Alexandrina. Retrieved 2022-04-30, from https://github.com/arcalex/linkgate. [URL Memento: archive.today]
dados.gov.pt. (n.d.). dados.gov.pt—Portal de dados abertos da Administração Pública [Website]. dados.gov.pt. Retrieved 2021-12-20, from https://dados.gov.pt/pt/. [URL Memento: Wayback Machine]
Deutsches Literaturarchiv Marbach. (n.d.). Deutsches Literaturarchiv Marbach [Website]. Deutsches Literaturarchiv Marbach. Retrieved 2021-12-19, from https://www.dla-marbach.de/?r=1. [URL Memento: Wayback Machine]
Digital Preservation Coalition. (n.d.). Digital Preservation Coalition [Website]. Digital Preservation Coalition. Retrieved 2021-10-08, from https://www.dpconline.org. [URL Memento: Wayback Machine]
Digital Preservation Coalition. (n.d.). Novice to Know-How—Digital Preservation Coalition [Web page]. Digital Preservation Coalition. Retrieved 2021-12-16, from https://www.dpconline.org/digipres/train-your-staff/n2kh-online-training. [URL Memento: Wayback Machine]
Egense, T. (2017). SolrWayback [GitHub]. NetarchiveSuite (netarchivesuite/solrwayback). Retrieved 2021-07-06, from https://github.com/netarchivesuite/solrwayback. [URL Memento: Wayback Machine]
Gephi. (n.d.). Gephi—The Open Graph Viz Platform [Website]. Gephi.Org. Retrieved 2021-08-12, from https://gephi.org/. [URL Memento: Wayback Machine]
Internet Archive. (2011). Heritrix [GitHub]. Internet Archive (internetarchive/heritrix3). Retrieved 2021-11-23, from https://github.com/internetarchive/heritrix3. [URL Memento: Wayback Machine]
International Internet Preservation Consortium. (2022). IIPC Member Organisations Google Map | IIPC [Google map]. Google My Maps. Retrieved 2022-05-12, from https://www.google.com/maps/d/viewer?mid=14Pe_dyH97jAKh1kBJZ0j5mfzGVY. [URL Memento: Wayback Machine]
International Internet Preservation Consortium. (n.d.). About Us | IIPC [Web page]. International Internet Preservation Consortium. Retrieved 2021-04-23, from https://netpreserve.org/about-us/. [URL Memento: Wayback Machine]
International Internet Preservation Consortium. (n.d.). Bibliography | IIPC [Web page]. International Internet Preservation Consortium. Retrieved 2021-09-09, from https://netpreserve.org/web-archiving/bibliography/. [URL Memento: Wayback Machine]
International Internet Preservation Consortium. (n.d.). Collection development policies | IIPC [Web page]. International Internet Preservation Consortium. Retrieved 2021-09-08, from https://netpreserve.org/web-archiving/collection-development-policies/. [URL Memento: Wayback Machine]
International Internet Preservation Consortium. (n.d.). IIPC Blog [Blog site]. IIPC Blog. Retrieved 2021-10-06,from https://netpreserveblog.wordpress.com. [URL Memento: Wayback Machine]
International Internet Preservation Consortium. (n.d.). IIPC Members | IIPC [Web page]. International Internet Preservation Consortium. Retrieved 2021-11-23, from https://netpreserve.org/about-us/members/. [URL Memento: Wayback Machine]

International Internet Preservation Consortium. (n.d.). IIPC TSS Webinar: Under the Hood of Solrwayback 4 | IIPC [Web page]. International Internet Preservation Consortium. Retrieved 2021-11-04, from https://netpreserve.org/events/iipc-tss-webinar-solrwayback4/. [URL Memento: Wayback Machine]

International Internet Preservation Consortium. (n.d.). LinkGate: Core Functionality and Future Use Cases [Web page]. International Internet Preservation Consortium. Retrieved 2021-09-10, from https://netpreserve.org/projects/linkgate/. [URL Memento: Wayback Machine]

Jupyter Team. (2015). The Jupyter Notebook—Jupyter Notebook 6.4.12 documentation [Wiki]. Jupyter Notebook/ReadTheDocs.Io. Retrieved 2021-10-07, from https://jupyter-notebook.readthedocs.io/en/stable/. [URL Memento: Wayback Machine]

Koninklijke Bibliotheek. (n.d.). Koninklijke Bibliotheek | KB, National Library of the Netherlands [Website]. Koninklijke Bibliotheek. Retrieved 2021-10-20, from https://www.kb.nl/en. [URL Memento: Wayback Machine]

Microsoft. (n.d.). Data Visualization | Microsoft Power BI [Web page]. Microsoft. Retrieved 2021-10-19, from https://powerbi.microsoft.com/en-us/. [URL Memento: Wayback Machine]

Microsoft. (n.d.). Microsoft Excel Spreadsheet Software | Microsoft 365 [Web page]. Microsoft. Retrieved 2021-05-24, from https://www.microsoft.com/en-ie/microsoft-365/excel. [URL Memento: Wayback Machine]

Milligan, I., Fritz, S., Ruest, N., & Lin, J. (2021). Building Community through Archives Unleashed Datathons: Lessons Learned [Conference presentation]. I*IPC General Assembly and Web Archiving Conference, June 14-16, 2021, Online*. Retrieved 2022-06-29, from https://digital.library.unt.edu/ark:/67531/metadc1827558/. [URL Memento: Wayback Machine]

NDSR Art. (2016, April 27). About | NDSR Art. National Digital Stewardship Residency (NDSR). Retrieved 2021-09-19, from http://ndsr-pma.arlisna.org/about/. [URL Memento: Wayback Machine]

NetLab. (n.d.). NetLab – Research Infrastructure Project [Website]. NetLab. Retrieved 2022-01-17, from https://www.netlab.dk. [URL Memento: Wayback Machine]

Olia Lialina, & Espenschie, D. (n.d.). About | One Terabyte of Kilobyte Age [Blog post]. One Terabyte of Kilobyte Age. Retrieved 2021-10-31, from https://blog.geocities.institute/about. [URL Memento: Wayback Machine]

pandas. (n.d.). pandas—Python Data Analysis Library [Website]. Pandas.Pydata. Retrieved 2022-04-20, from https://pandas.pydata.org. [URL Memento: Wayback Machine]

Penn Library. (n.d.). Lib Guide: Web Archiving for the Arts and Historic Preservation [Web page]. Penn Libraries. Retrieved 2022-06-27, from https://guides.library.upenn.edu/fisherwebarchive/home

RESAW. (n.d.). About RESAW [Web page]. Research Infrastructure for the Study of Archived Web Materials (RESAW). Retrieved 2021-03-10, from http://resaw.eu/about/. [URL Memento: Wayback Machine]

Rhizome. (n.d.). Conifer (previously Webrecorder) [Website]. Conifer. Retrieved 2021-05-19, from https://conifer.rhizome.org. [URL Memento: Wayback Machine]

Rhizome. (n.d.). Conifer | About [Web page]. Rhizome. Retrieved 2021-04-23, from https://conifer.rhizome.org/_faq. [URL Memento: Wayback Machine]

Sherratt, T. & Andrew Jackson. (2021). GLAM Workbench—Web Archives [Wiki]. GLAM Workbench/GitHub.Io. Retrieved 2022-04-29, from https://glam-workbench.github.io/web-archives. [URL Memento: Wayback Machine]
Tableau. (n.d.). Tableau: We're changing the way you think about data [Website]. Tableau. Retrieved 2021-10-26, from https://www.tableau.com/en-gb. [URL Memento: Wayback Machine]
TARA. (n.d.). TARA [Web page]. Trinity College Dublin. Retrieved 2021-11-16, from http://www.tara.tcd.ie. [URL Memento: Wayback Machine]
The Archives Unleashed Project. (n.d.). The Archives Unleashed Project [Website]. The Archives Unleashed Project. Retrieved 2021-11-23, from https://archivesunleashed.org. [URL Memento: Wayback Machine]
The Archives Unleashed Project. (n.d.). The Archives Unleashed Toolkit [Web page]. The Archives Unleashed Project. Retrieved 2021-10-21, from https://archivesunleashed.org/aut. [URL Memento: Wayback Machine]
The National Archives. (n.d.). The National Archives [Website]. The National Archives. Retrieved 2021-10-07, from https://www.nationalarchives.gov.uk. [URL Memento: Wayback Machine]
Trinity College Dublin. (n.d). Trinity College Dublin, the University of Dublin, Ireland [Website]. Trinity College Dublin. Retrieved 2021-12-20, from https://www.tcd.ie. [URL Memento: Wayback Machine]
WARCnet. (n.d.). About WARCnet | WARCnet [Web page]. Aarhus University. Retrieved 2012-08-29, from https://cc.au.dk/en/warcnet/about/. [URL Memento: Wayback Machine]
Waring, A. (2012, August 29). Multimodal Methods for Analysing Communication and Learning with Digital Technologies – MODE Summer School: 24-28th June 2013 [Web page]. MODE: Multimodal Methodologies, UCL Institute of Education. Retrieved 2012-08-29, from https://mode.ioe.ac.uk/2012/08/29/analysing-digital-data-and-environments-mode-summer-school-24-28th-june-2013/. [URL Memento: Wayback Machine]
Webrecorder. (n.d.). ArchiveWeb.page [Website]. ArchiveWeb.Page. Retrieved 2022-01-05, from https://archiveweb.page. [URL Memento: Wayback Machine]
Zenodo. (n.d.). Zenodo—Research [Website]. Zenodo. Retrieved 2021-12-22, from https://zenodo.org/. [URL Memento: Wayback Machine]

# Appendix A: Information sheet

**Information Sheet, Web Archives - Researcher Skills & Tools Survey**

***Introduction***

Thank you for taking the time to consider participating in this survey.

Web Archives - Researcher Skills & Tools Survey is a collaborative research study. The study will be carried out by researchers from Maynooth University and the British Library. The project research will be led by Sharon Healy and supervised by Dr Joseph Timoney (Department of Computer Science, Maynooth University) and Prof Jane Winters (School of Advanced Study, University of London). The findings and results will be published as part of the WARCnet Papers. Data will also be used to inform the PhD dissertation of Sharon Healy, and future publications related to this. Sharon Healy and Dr Joseph Timoney will act as the data controllers for the collection, management, and storage of the data.

This study has been reviewed and received ethical approval from Maynooth University Research Ethics committee (SRESC-2021-2436150).

This is an anonymous survey and will take approximately 15 minutes to fill out. You may exit at any time during the process of filling out this survey, and your responses will not be recorded. If you wish to participate, simply complete the survey and click on submit, and your responses will be recorded as anonymous. If you decide to participate, it is important that you fully understand what is required. Please click next to read more information about the requirements and how the data will be collected and managed. Please note, it is equally important to attain participation from respondents who are novice users, as it is to attain responses from regular or experienced users.

***Purpose of the Project***

This survey study seeks to identify, and document skills and knowledge required to achieve a range of different research goals within web archiving. It will investigate skills that are useful or important for conducting research with web archives (develop a skills matrix); and the availability of resources to train or inform researchers of how to acquire these skills (list of resources). This study will investigate the methodological, technical, and legal challenges for using web archives for research; and will provide insights, to inform future investigations of potential solutions.

***What's Involved?***

What do you have to do?

You must be 18 years of age or over. If you decide to take part, you will be required to complete a questionnaire consisting of 28 questions, first on some basic demographic information and then some questions on your use of web archives.

***How will the information collected by this survey be used?***

The findings and results will be published as part of the WARCnet Papers. Data will also be used to inform the PhD dissertation of Sharon Healy, and future publications related to this. Sharon Healy is a

PhD Candidate and GOIPG Irish in Digital Humanities in the Department of Computer Science, Maynooth University. Opinions and data will be reported in an aggregated form. Any quotations from the data will be used in a manner that does not identify a participant. Sharon Healy will act as the data controller for the collection, management, and storage of the data.

***Who will have access to this data?***

This data will not be shared with a third party. The data will only be shared between the named researchers responsible for conducting the research, and the named data controller responsible for the long-term preservation of the data. The research will only be processed in a manner compatible with the purposes of this research, by the researchers concerned. Sharon Healy will act as the data controller for all responses and information gathered and will endeavour to store and preserve this data for a period of ten years as outlined in Maynooth University Research Integrity Policy.

(Please Note: It must be recognized that, in some circumstances, confidentiality of research data and records may be overridden by courts in the event of litigation or in the course of investigation by lawful authority. In such circumstances the University will take all reasonable steps within law to ensure that confidentiality is maintained to the greatest possible extent.)

***What if there is a problem?***

If you have any concerns or would like any further information about this research study, please contact Sharon Healy (sharon.healy@mu.ie), or the supervisor of this research Dr Joseph Timoney.

**INFORMED CONSENT**

By clicking the Boxes below, and submitting this survey, you are also confirming that:

☐ you are 18 years of age or over
☐ you have been sufficiently informed about the research study
☐ you understand the limits of confidentiality as described in the information sheet
☐ you are taking part in this research study voluntarily
☐ you understand that you can withdraw from the study while participating, and your responses will not be recorded
☐ you agree to have your responses stored, processed, and preserved in a manner compatible with the purposes of this research
☐ you agree to have your responses stored, processed, and preserved in a manner compatible with the purposes of this research

Permissions for Publication

I understand that my data, in an anonymous format, may be used if I give permission below:

☐ I agree to quotation/publication of extracts of data I provide
☐ I do not agree to quotation/publication of extracts of data I provide

***Other Information***

If during your participation in this study you feel the information and guidelines that you were given have been neglected or disregarded in any way, or if you are unhappy about the process, please contact the Secretary of the Maynooth University Ethics Committee at research.ethics@mu.ie or +353 (0)1 708 6019. Please be assured that your concerns will be dealt with in a sensitive manner.

For your information the Data Controller for this research project is Maynooth University, Maynooth, Co. Kildare. Maynooth University Data Protection officer is Ann McKeon in Humanity house, room 17, who can be contacted at ann.mckeon@mu.ie. Maynooth University Data Privacy policies can be found at https://www.maynoothuniversity.ie/data-protection.

***Custom Thank You***

Thank You for participating in this research, by filling out this survey. Please feel free to forward the link to this survey to colleagues in cultural heritage organisations and academic institutions. (https://www.onlinesurvey.com-link)

The results from this survey are anonymised. However, if you would like to be contacted at some stage in the future for focus groups on using web archives and archived web content, please email Sharon Healy (sharon.healy@mu.ie) with your name and position. Please note that providing this information does not compromise the confidentiality and anonymity of the survey. It is impossible to link an email sent to this address to a survey response.

If you would like further information about this research or if you have concerns/questions you would like to discuss about the research, please contact the principal researcher:

Sharon Healy: (sharon.healy@mu.ie) PhD Candidate & GOIPG IRC Scholar in Digital Humanities, Maynooth University (ORCID iD: https://orcid.org/0000-0003-3493-0938)

# Appendix B: Survey questions

**Survey Questions, Web Archives - Researcher Skills & Tools Survey**

Part 1 - About You

**DEMOGRAPHICS**

These questions allow for the exploration of any trends from the rest of the survey across nationality, age, gender, position, and research interests.

denotes a Required field **

**Q.1** - What is your current country of residence? **

*Dropdown Box - Country Index*

**Q.2** - Please select your age? **

*Multiple Choice*

18-24 25-34 35-44 45-54 55-64 65+ Prefer not to say

**Q.3** - What gender do you identify with? **

*Multiple Choice*

Male Female Other Prefer not to say

**Q.4** – Please describe your position? **

(e.g. PhD student in Sociology; Web archivist; IT specialist in a library; Senior lecturer in Media Studies; Retired historian; Unemployed researcher)

*Text Box*

**Q.5** – Please describe in your own words your research interests in general? **

*Text Box*

**Part 2 - Types of Data & Tools**

The following questions relate to the kinds of data used in your research, your research outputs, and the types of tools you use for conducting your research with web archives.

**Q.6** – What type of data do you collect as part of your research in working with web archives and archived web content? **

*Tick Boxes*

- WARC files
- Text files
- Audio files
- PDF files
- Screenshots
- Images (eg. photographs)
- GIFs
- Button Icons
- Banners
- Numerical data (e.g. statistics)
- HTML code
- URLs
- Crawl logs
- Tracking cookies
- Archival metadata
- Other - please specify

**Q.7** – What type of tools do you use to COLLECT your data? - please list all tools that apply

*Text Box*

**Q.8** – What type of tools do you use to ANALYSE your data? - please list all tools that apply

*Text Box*

**Q.9** – What type of data do you output as part of your research in working with web archives, e.g. spreadsheet, screenshot, text fragment etc. - please list all that apply

*Text Box*

**Part 3 - Skills & Knowledge**

This section looks at the skills and knowledge of researchers for conducting research with web archives.

**Q.10** – Please describe in your own words your primary areas of research/curation with web archives? **

*Text Box*

**Q.11** – What led you to using web archives for your research? **

*Text Box*

**Q.12** – How long have you been using web archives for your research? **

*Multiple-Choice*

- 0-6 months
- 6 months - 1 year
- 1-2 years
- 3-5 years
- 5-10 years
- 10-15 years
- More than 15 years

**Q.13** – What web archive(s) do you use for your research? - please tick all that apply **

*Multiple-Choice*

- Archive.today, http://archive.is/
- Arquivo.pt (FCT | FCCN, Portugal), https://arquivo.pt/
- BnF Archives de l'internet (Bibliothèque nationale de France), https://www.bnf.fr/fr/archives-de-linternet
- Common Crawl, https://commoncrawl.org/

- Government of Canada Web Archive, https://www.bac-lac.gc.ca/eng/discover/archives-web-government/Pages/web-archives.aspx
- INA Web Archive (Institut Nationale de l'Audiovisuel), https://institut.ina.fr/collections/le-web-media
- Internet Archive, Wayback Machine, http://archive.org/web/
- Luxembourg Web Archive (Bibliothèque Nationale de Luxembourg) https://bnl.public.lu/fr/rechercher/outils-recherche/webarchive.html
- Netarkivet, Denmark (the Royal Library, and the State and University Library), http://netarkivet.dk/
- NLI Web Archive (National Library of Ireland), https://archive-it.org/home/nli
- PRONI Web Archive (Public Records Office of Northern Ireland), https://www.nidirect.gov.uk/services/search-proni-web-archive
- Time Travel, http://timetravel.mementoweb.org/
- UK Web Archive (British Library), https://www.webarchive.org.uk/ukwa/
- UK Government Web Archive (UK National Archives), http://www.nationalarchives.gov.uk/webarchive/
- UK Parliament Web Archive (UK Parliament), http://webarchive.parliament.uk/
- US Library of Congress Web Archive, https://www.loc.gov/websites/collections/
- Webarchief van Nederland (Koninklijke Bibliotheek),http://www.kb.nl
- Other - please specify

**Q.14** – What barriers did you encounter when working with web archives and how did you overcome (or workaround) them? **

*Text Box*

**Q.15** – What skills or knowledge did you have BEFORE starting your research in web archives that proved useful? Please tick all that apply **

*Likert Scale*

| TOPIC | No - I had NO knowledge | Yes - I had SOME knowledge | Yes - I had a LOT of knowledge |
|---|---|---|---|
| How websites are built/ made/ updated | X | X | X |
| How the internet works - Geo-IP, servers, browsers, domains, hosting etc. | X | X | X |
| How web archiving works - WARCs, Capture tools, storage, and playback | X | X | X |
| How digital curation works - collection, metadata, storage, access, long-term preservation | X | X | X |
| How Fair Use works - copyright, reproduction rights, fair use | X | X | X |
| How digital legal deposit works and what it is | X | X | X |
| Excel (or other spreadsheet) - Intermediate/Advanced | X | X | X |
| Data analysis, such as topic modelling, textual analysis, etc. | X | X | X |

| | | | |
|---|---|---|---|
| Metadata analysis | X | X | X |
| Database creation and maintenance | X | X | X |
| Python - Basic/intermediate | X | X | X |
| Java - Basic/intermediate | X | X | X |
| httrack | X | X | X |
| Other - please specify: | | | |

**Q.16** – What skills or knowledge do you WISH you had before you started your research in web archives? please tick all that apply **

*Likert Scale*

| TOPIC | No Opinion | Yes - I wish I had SOME knowledge about this before I started my research | Yes - I wish I had a LOT of knowledge about this before I started my research |
|---|---|---|---|
| How websites are built/ made/ updated | X | X | X |
| How the internet works - Geo-IP, servers, browsers, domains, hosting etc. | X | X | X |
| How web archiving works - WARCs, Capture tools, storage, and playback | X | X | X |
| How digital curation works - collection, metadata, storage, access, long-term preservation | X | X | X |
| How Fair Use works - copyright, reproduction rights, fair use | X | X | X |
| How digital legal deposit works and what it is | X | X | X |
| Excel (or other spreadsheet) - Intermediate/Advanced | X | X | X |
| Data analysis, such as topic modelling, textual analysis, etc. | X | X | X |
| Metadata analysis | X | X | X |
| Database creation and maintenance | X | X | X |
| Python - Basic/intermediate | X | X | X |
| Java - Basic/intermediate | X | X | X |
| httrack | X | X | X |
| Other - please specify: | | | |

**Q.17** – What new skills did you learn AFTER starting your research in web archives? please list all that applies

*Text Box*

**Q.18** – Did your research question or parameters change AFTER starting your research project? ** (including the disruptions caused by the COVID pandemic)

*Multiple choice*

- Yes - they changed a lot
- Yes - they changed a little
- No - they did not change

**Q.19** – If so, how? If you answered Yes to the question above, please describe how your research question or parameters changed AFTER starting your research project

*Text Box*

**Part 4 - Data Citation**

This section looks at the citation systems you use for conducting research with web archives.

**Q.20** – What standard of referencing system do you use for citing sources in your research in general?

*Tick Boxes*

- MLA (Modern Languages Association) system
- APA (American Psychological Association) system
- Harvard system
- MHRA (Modern Humanities Research Association) system
- IEEE (Institute of Electrical and Electronics Engineers) system
- Other - please specify (add comment box)

**Q.21** – Do you have any challenges when citing archived web content from a web archive?

*Yes / No / Sometimes, Checkboxes*

**Q.22** – If you answered Yes to the question above, could you please describe some of the challenges you have for citing archived web content?

*Text Box*

**Q.23** – Do you have any challenges when citing datasets of archived web content?

*Yes / No / Sometimes, Checkboxes*

**Q.24** – If you answered Yes to the question above, could you please describe some of the challenges you have for citing datasets of archived web content?

*Text Box*

**Part 5 - Resources**

This section looks at resources you found useful to further your skills and knowledge in your research with web archives.

**Q.25** – Please list any resources that were useful to you to further your skills and knowledge in your research with web archives. This could be an online or in person training course, workshop or mentorship?

*Text Box*

**Q.26** – Have you shared any data you collected or created in an institutional or subject repository?

*Yes / No, Multiple choice*

**Q.27** – If you answered Yes to the question above, please name the repository(s) where your data is stored/shared? Also, could you please provide a link to the repository

*Text Box*

**Q.28** – OPTIONAL: Any other comments you would like to add

*Text Box*

SUBMIT >>>>

Custom Thank You

Thank You for participating in this research, by filling out this survey. Please feel free to forward the link to this survey to colleagues in cultural heritage organisations and academic institutions. (https://www.surveymonkey.com/xxxx)

The results from this survey are anonymised. However, if you would like to be contacted at some stage in the future for focus groups on using web archives and archived web content, please email Sharon Healy (sharon.healy@mu.ie) with your name and position. Please note that providing this information does not compromise the confidentiality and anonymity of the survey. It is impossible to link an email sent to this address to a survey response.

If you would like further information about this research or if you have concerns/questions you would like to discuss about the research, please contact the following researcher: Sharon Healy: (sharon.healy@mu.ie) PhD Candidate & GOIPG IRC Scholar in Digital Humanities, Maynooth University (ORCID iD: https://orcid.org/0000-0003-3493-0938)

# Appendix C: Comparison for challenges encountered

Table C.1: Breakdown of combined thematic representations of participant responses for challenges encountered when working with web archives, by participants who identified with working in a Library, Archive or Web Archive environment (n=27), in line with novice, intermediate or experienced levels

| Combined thematic representations for challenges encountered by participants who identified with working in a Library, Archive or Web Archive environment (n=27) | Novice 0-2 years | Novice-Inter. 3-5 years | Inter. 5-10 years | Experienced 10-15/ +15 years |
|---|---|---|---|---|
| \|> Inconsistencies and Incompleteness (r=11) | | | | |
| ● Broken links to files | r=1 | | r=1 | r=1 |
| ● Erroneous/incomplete crawls | r=1 | r=2 | | |
| ● Layout/visual deficiencies | | r=1 | r=1 | |
| ● Capturing dynamic content | | | r=1 | |
| ● Inconsistency with crawl frequency of early websites | | | r=1 | |
| ● R: "Variation in what is collected over time" | | | | r=1 |
| \|> Legalities for acquisition/providing access (r=8) | | | | |
| ● Acquisition restrictions for selective archiving | r=1 | | | |
| ● Challenges to get permissions for selective archiving | r=1 | | | |
| ● Embargos | r=1 | | | |
| ● Challenges to provide access due to legal/Copyright/GDPR | | r=4 | | r=1 |
| \|> Technical challenges (r=8) | | | | |
| ● Data storage | r=1 | | | |
| ● Lack of IT infrastructure | r=1 | | | |
| ● Data processing | | r=1 | | |
| ● Search and discovery | | r=1 | | |

| | | | | |
|---|---|---|---|---|
| ● Challenges to save sites due to firewall/security | | | r=1 | |
| ● Difficult to create bulk data sets to share with researchers | | | | r=1 |
| ● File format obsolescence | | | | r=1 |
| ● Technical challenges (in general) | | | | r=1 |
| \|> Challenges with learning new skills (r=6) | | | | |
| ● R: "It was a bit strange at first because I didn't have much of an idea of web archiving since I was more used to working with paper" | r=1 | | | |
| ● R: "learning curve was steep" | r=1 | | | |
| ● R: "Limited technical skills to analyse the WARC-files and the information within them" | r=1 | | | |
| ● R: "complexity of the WARC files" | | r=1 | | |
| ● R: "Learning how to use research tools (from a non-technical user's perspective)" | | | | r=1 |
| ● R: "Need to learn a lot about what web archives are" | | | | r=1 |
| \|> Volume of data (r=2) | | | | |
| ● R: "scale of the archive" | | r=1 | | |
| ● R: "The size of the collections and the difficulty of narrowing down a set of data that is manageable and appropriate" | | | | r=1 |
| \|> Producing documentation/metadata (r=2) | | | | |
| ● R: "confusing records" | r=1 | | | |
| ● R: "Trying to guess the date when the site may have been crawled and when changes happen." | | | r=1 | |
| \|> Financial challenges (r=4) | r=2 | r=1 | | r=1 |
| ● Cost of services | r=1 | | | |
| ● Cost of storage | r=1 | | | |

| | | | | |
|---|---|---|---|---|
| ● Attaining funding | | r=1 | | |
| ● "On-premises access to web archives makes them economically inaccessible" | | | | r=1 |
| \|> Institutional challenges (r=1) | | | | |
| ● "a barrier can be institutional in convincing other areas of the organization about the value of the web archive" | | r=1 | | |
| \|> Conceptual challenges (r=1) | | | | |
| ● The main ones are conceptual | | | | r=1 |

Table C.2: Breakdown of combined thematic representations of participant responses for challenges encountered when working with web archives, by participants who identified with being a Scholar, Academic, Lecturer, Student, or IT/ Web Design environment (n=9), in line with novice, intermediate or experienced levels.

| Combined thematic representations for challenges encountered by participants who identified with Scholar, Academic, Lecturer, Student, or IT/Web Design environment (n=9) | Novice 0-2 years | Novice-Inter. 3-5 years | Inter. 5-10 years | Experienced 10-15/ +15 years |
|---|---|---|---|---|
| \|> Inconsistencies and Incompleteness (r=10) | | | | |
| ● Inconsistencies in terms of what was saved | r=1 | r=2 | r=3 | |
| ● Inconsistent temporal coverage | | r=1 | r=1 | |
| ● Incompleteness in the data itself | | | r=1 | |
| ● Layout/visual deficiencies | | | r=1 | |
| \|> Challenges in an IT/ Business/ Administrative environment (r=2) | | | | |
| ● R: "Dependency on a not-for-profit, third-party archiving initiative to meet our business needs " | r=1 | | | |
| ● Funding and low awareness from stakeholders | | | r=1 | |
| \|> Challenges with learning new skills (r=6) | | | | |
| ● Challenges with tools for web archives research | | r=1 | | |

| | | | | |
|---|---|---|---|---|
| ● Difficulties to understand how web archives are set up | | r=1 | | |
| ● Having to acquire new programming skills | | r=1 | | r=1 |
| ● Learning about the limitations of replay interfaces | | | r=1 | |
| ● Learning what a WARC file was | | | r=1 | |
| \|> Legalities on access, use, and storage (r=8) | | | | |
| ● Legal challenges regarding access to data | | r=2 | r=2 | r=1 |
| ● Inability to download data | | r=1 | | |
| ● Legal challenges regarding use of data | | | | r=1 |
| ● Legal challenges regarding storage of data | | | | r=1 |
| \|> Performance related issues (r=1) | r=1 | | | |
| \|> Research methods and approaches (r=5) | | | | |
| ● Combining traditional methods with web archives research | | r=1 | | |
| ● Lack of research methods/theory | | r=2 | | |
| ● Data analysis | | | r=1 | |
| ● Archived web as a source for research | | | | r=1 |
| \|> Lack of documentation/metadata (r=2) | | | | |
| ● R: “lack of of archival context” | | | r=1 | |
| ● R: "issues relating to the lack of documentation" | | | | r=1 |
| \|> Volume of data for research(r=2) | | | | |
| ● R: “volume” | | | r=1 | |
| ● R: “Working with large-scale data” | | | | r=1 |

# RESUMÉ

Denne undersøgelse er en del af et samarbejdsprojekt mellem forskere fra Maynooth University, British Library, International Internet Preservation Consortium (IIPC), Statsbiblioteket i Bayern og Siegen Universitet. Forskerne er alle medlemmer af forskningsnetværket Web ARChive, der undersøger webdomæner og begivenheder (WARCnet, warcnet.eu). WARCnet finansieres af Danmarks Frie Forskningsfond | Kultur og Kommunikation (grant no 9055-00005B).

Undersøgelsen fokuserer på personer fra hele verden, der arbejder med i webarkiver i forbindelse med webarkivering, kuratering og brug af webarkiver og arkiveret webindhold til forskning eller til andre formål. Den er som sådan rettet mod både skabere og brugere af webarkiver. Vi anser webarkiveringsarbejde for at være repræsentativt for de processer og aktiviteter, der er beskrevet i Archive-Its livscyklusmodel for webarkivering, fra vurdering, accession og bevaring, til genafspilning, adgang, brug og genbrug (Bragg & Hannah, 2013). Undersøgelsen forsøgte at identificere og dokumentere de færdigheder, værktøjer og den viden, der kræves for at opnå en bred vifte af mål inden for webarkiveringslivscyklussen, og for at udforske udfordringerne for deltagelse i webarkivforskning samt mellemrummene imellem sådanne udfordringer på tværs af praksisfællesskaber. Vi konkluderer, at der er et løbende behov for at undersøge færdigheder, værktøjer og metoder forbundet med webarkiveringslivscyklussen, så længe internet-, web- og softwareteknologier bliver ved med at udvikle sig.

Undersøgelsesmetoderne omfattede skrivebordsforskning, deltagelse i WARCnet-mødediskussioner og et online spørgeskema. Spørgeskemaet blev rundsendt via sociale medier og e-mail fra 23. juli til 21. september 2021. Strategien var at rekruttere målrettet blandt arkivarer, bibliotekarer, kuratorer, informationschefer, forskere, studerende, historikere osv. og bestod af opslag på sociale medier og rekrutteringsmails til netværkslister for arkivarer, bibliotekarer, kuratorer, digital humaniora, internetstudier, og webarkivforskning. Resultaterne er baseret på et endeligt antal svar fra 44 deltagere.

## Demografi

I denne undersøgelse er deltagerne (N=44) i alderen mellem 18 og 64 år og beskriver sig selv som bosiddende i Nordamerika, Europa og Asien. Deltagerne beskriver sig selv som værende på begynder-, mellem- og erfarent niveau i forhold til at arbejde med eller bruge webarkiver, og der er en ligelig fordeling mellem deltagere, der identificerer som mand og kvinde. Dette

kan være et tegn på, at køn ikke udgør en åbenlys barriere i webarkivforskning, i hvert fald i denne undersøgelse.

Med hensyn til deltagernes positionelle baggrund har vi opdelt dem i to tematiske grupper, nemlig (i) deltagere, der identificerede sig med at arbejde i et biblioteks-, arkiv- eller webarkivmiljø (n=30), og (ii) deltagere, der identificerede sig som værende forsker, akademiker, underviser, ph.d.-studerende eller ansat i IT (n=14). I første omgang troede vi, at det ville være muligt at afstemme deltagernes holdninger med, om de var skabere af webarkiver eller brugere af webarkiver, men det var ikke tilfældet. Faktisk blev grænserne noget slørede, da nogle respondenter i webarkiveringsfællesskabet også angiver, at de er brugere af webarkiver som en del af deres arbejde; mens nogle respondenter fra forskningsmiljøerne angiver, at de er skabere/kuratorer af webarkiver til forskningsformål. Kategoriseringen af deltagernes holdninger var således ikke så entydig som vi oprindeligt forestillede os, og vi erkender, at der er et vist overlap.

På baggrund af deltagernes interesser, baggrund, erfaringer og deres forhold til webarkivforskning har vi tentativt inddelt dem i grupper med tilknytning til et eller flere af følgende emneområder, i alfabetisk rækkefølge:

- Arts, Humanities, Digital Humanities, Social Sciences, Media Studies (Kunst, Humaniora, Digital Humaniora, Samfundsvidenskab, Medievidenskab)
- Brug af webarkiver og arkiveret webindhold
- Business og/eller Jura
- Datavidenskab/-analyse, Statistik
- Informationsvidenskab (undtagen webarkivering/kuratering)
- Internet/web applikationer, systemer
- IT-drift/Computerapplikationer, -systemer, -miljøer
- Webarkiver, webarkivering, kuratering

# Hovedresultater og Indsigter

I nærværende opsummering giver vi et overblik over resultater, konklusion og diskussion og inddelt i følgende fire hovedafsnit:

- Færdigheder, viden, værktøjer og metoder i webarkivforskning
- Udfordringer ved webarkivforskning
- Udfordringer ved pligtaflevering, ophavsret og GDPR
- Vigtigheden af samarbejde

## Færdigheder, viden, værktøjer og metoder i webarkivforskning

Ud fra resultaterne har vi fremsat en bred vifte af færdigheder, værktøjer, metoder og viden, som er nødvendige, ønskværdige eller nyttige for webarkivforskningsdomænet på tværs af praksisfællesskaber. Blandt de vigtigste er:

- Software og værktøjer
- Webarkiver, webarkivering, kuratering
- Programmering, scripting-sprog
- Digitale kurateringsprocesser/-arbejdsgange
- Data-analysefærdigheder
- Undersøgelsesmetoder/-tilgange
- Web design/internetrelaterede færdigheder
- Informationsvidenskab (undtagen webarkivering/kuratering)

Undersøgelsen viser flere fælles træk mellem deltagere, der beskrev sig selv som arbejdende i et biblioteks-, arkiv- eller webarkivmiljø, og deltagere, der identificerede sig som værende forskere, akademikere, undervisere, studerende eller arbejdende i et IT-/webdesignmiljø. For eksempel angiver respondenter fra begge grupper, at de benytter webarkiver til at finde information, litteratur og gamle websteder og udviser enslydende bekymringer om datatab og ændringer i webindhold. Håndtering af exceptionelt store datamængder nævnes yderligere som en udfordring af respondenter fra begge grupper. Og respondenter fra begge grupper indikerer vigtigheden af at tilegne sig viden, samt tekniske og kritiske færdigheder gennem træning, kurser og workshops samt gennem samarbejder og mentorskab. Det fremgår af forskellige dele af resultaterne, at flere respondenter fra begge grupper indikerer, at der er behovet for samarbejde og veje til at udvikle yderligere forbindelser mellem skabere/kuratorer og brugere/forskere.

Med hensyn til værktøjer og metoder vil begge grupper drage fordel af træning i forskellige indsamlingsmetoder, herunder crawl-software, skærmbillede-, skærmoptagelses- og screencasting-værktøjer og værktøjer til at downloade data fra API'er. Der er også tegn på, at udvikling af undervisningsmateriale i brugen af regnearkssoftware og håndtering og bevaring af regneark som dataoutput vil være nyttigt for begyndere og øvede og på mere avancerede niveauer på tværs af webarkivforskningskredse som helhed. Undersøgelsen indikerer desuden, at brugere af webarkiver vil have gavn af introducerende kurser for webarkivering, mens personale i et webarkiveringsmiljø vil have gavn af at opnå en vis forståelse og træning i de værktøjer og metoder, som brugere/forskere anvender til at analysere arkiverede webdata. Det bør dog nævnes, at undersøgelsen viser, at deltagere fra videnskabelige eller akademisk miljøer benytter sig af en mangfoldighed af værktøjer og metoder. Desuden har

forskningsspørgsmål eller –metode ofte en indflydelse på, hvilke værktøjer og metoder der vælges, fx i tilfælde hvor data indsamles manuelt til nærstudier, eller når kun bestemte dele af en hjemmeside nedtages. Denne gruppe af deltagere står også over for udfordringer på grund af mangel på forskningsmetoder, teori og tilgange til at kombinere traditionelle metoder med webarkivforskning. Begge grupper vil således drage fordel af fælles træning i form af aktuelle forskningstilgange og metoder til brug af arkiveret web, inklusive demonstrationer af værktøjer og software. På denne måde ville området blive beriget gennem input fra en dialog med begge grupper ved udviklingen af en bedre forståelse for forskningsmetoder og tilgange til brug af webarkiver, samt for "Tilegnelse af en ordentlig forståelse for arkiveret web som en specifik kildetype og konsekvenserne af disse karakteristika" til forskning ved hjælp af det arkiverede web, som påpeget af en respondent.

## Udfordringer ved webarkivforskning

Undersøgelsen identificerer flere udfordringer, der påvirker på tværs af praksisgrupperne. For eksempel resulterer udfordringer med at fange dynamisk webindhold ofte i mangler i arkivet, hvilket yderligere kan vise sig for slutbrugeren som usammenhængende og ufuldstændige arkivkopier. Ufuldstændighed på grund af manglende elementer eller ødelagte links på live-websteder er problematisk for både webarkivarer og slutbrugere, især når hullerne er svære at dokumentere og forklare for brugerne. Produktion af omfattende metadata og dokumentation til webarkivsamlinger er en enorm udfordring for de institutioner, der fremstiller arkiverne, da det er en tidskrævende og arbejdskrævende proces, der forværres af det enorme dataomfang. Disse ufuldstændige metadata og dokumentation kommer så til at give problemer for slutbrugeren, der søger at arbejde med samlingerne. Derudover kan mangel på ressourcer og specialiserede kompetencer også hæmme udviklingen af omfattende dokumentation, hvilket ellers kunne øge mangfoldigheden af brugere, som yderligere har forskellige niveauer af færdigheder og erfaring. Der er også behov for at overveje, at akademiske forskere og andre slutbrugere såsom journalister eller advokater måske ikke kan dedikere tid eller energi til at opnå en god forståelse af disse problemstillinger, og det kan derfor opfattes som en adgangsbarriere eller hindring for at beskæftige sig med webarkiver. Der ville således være en vis fordel ved at give brugere og potentielle brugere en introduktion til webarkivering i et omfang, der passer til det webarkiv, der anvendes, i et forsøg på at øge fokus og derigennem give en større forståelse for samlingernes afgrænsninger for så vidt angår arkiveringsstrategiernes begrænsninger på grund af tekniske udfordringer, juridiske begrænsninger og mangel på ressourcer. Det giver også mulighed for samarbejde mellem webarkiver og deres brugere for at udvikle dokumentation sammen, hvilket i sidste ende kunne skræddersyes på tværs af discipliner og professioner. Dette ville

være en betydelig gevinst for begge grupper og skabe en god cyklus mellem arkivskabelse og -anvendelse.

Udfordringer med at lære nye færdigheder opleves af respondenter fra begge grupper. Vi fremhæver, hvordan begge grupper kunne drage fordel af muligheden for fælles uddannelse på tværs af hele spektret af aktiviteter i webarkiveringslivscyklussen. Undersøgelsen giver et overblik over, hvilke typer af færdigheder og viden webarkivarer og webarkivbrugere havde forud for arbejdet med webarkiver, de færdigheder de udviklede under arbejdet med webarkiver og de udfordringer de stod over for med denne type af kilde. Vi foreslår, at dette kan bruges som et udgangspunkt for at fremme en diskussion om at udvikle effektivt uddannelsesmateriale til at opnå de nødvendige færdigheder og værktøjer til at arbejde med webarkiver på tværs af spektret af arkivskaber, kurator, tekniker eller bruger/forsker. Vi foreslår endvidere, at sådan uddannelse også skal benchmarkes i en "færdighedsmatrice", da det er meget svært at udvikle og give tilstrækkelig uddannelse uden et benchmark at måle i forhold til. Vi har også erfaret, at de udfordringer, som deltagerne i undersøgelsen oplever, ikke bliver mindre selv med stigende erfaring, og vi fremhæver derfor behovet for uddannelseuanset færdighedsniveau. Vi foreslår, at der er behov for yderligere forskning for at udvikle målrettede læringsmateriale til både introduktion og mere avanceret uddannelse, således at man kan se, hvordan udfordringerne veksler i forhold til erfaringsniveau tværs af grupper.

## Udfordringer ved pligtaflevering, ophavsret og GDPR

Juridiske udfordringer relateret til fx pligtaflevering, ophavsret og GDPR udgør barrierer for både webarkivering og forsker-/brugergrupper. Respondenter fra begge grupper diskuterer udfordringer ved at henvise til arkiveret webindhold fra pligtafleverede arkiver eller arkiver med begrænset adgang. Deltagere, der arbejder med webarkiveringsgruppen, nævner udfordringer med at give adgang til arkiverede websamlinger på grund af lovgivning, ophavsret, GDPR og klausuleringer. Udfordringer på grund af lave svarprocenter med at opnå tilladelser fra webstedsejere er også nævnt, både for indsamling af websteder og i relation til at give adgang til de arkiverede websteder uden for et læsesalsmiljø. Yderligere fremhævet er det forhold, at selvom pligtaflevering muliggør indsamling af websteder af en pligtafleveringsinstitution, beskæftiger sådanne regler sig ofte ikke effektivt med adgangsmuligheder. For nogle institutioner kan der kun gives adgang på stedet, hvilket "gør dem økonomisk utilgængelige", som en af respondenterne bemærker. Dette forhold bør undersøges målrettet, da der har været meget lidt opmærksomhed på de socioøkonomiske faktorer, som kan hindre adgang til og arbejde med webarkiver.

Deltagere, der placerede sig selv i den akademiske gruppe, diskuterer udfordringer ved brugen af webarkiver på grund af juridiske kravrelateret til adgang til data, brug af data og opbevaring af data fra webarkiver. Andre udfordringer omfatter håndtering af ophavsretligt beskyttede data fra et webarkiv, samt manglende mulighed for at downloade data fra nogle webarkiver. Der er også udfordringer ved internationale samarbejdsprojekter på grund af forskellig lovgivning om pligtaflevering på tværs af forskellige lande, som påvirker, hvordan data kan tilgås og anvendes og af hvem. Derudover er disse udfordringer med at dele data fra webarkiver eller gøre dem genanvendelige i modstrid med aktuelle tendenser fra fonde og andre finansieringskilder, som i stigende grad stiller krav om åben adgang og åbne videnskabelige rammer for forskning og dataoutput. Vi foreslår, at yderligere diskussion og samarbejde er påkrævet for at udvikle håndteringen forskningsdata inden for rammerne af pligtafleveringslovgivning, open science og forskningsmiljøer i webarkivering. Som udgangspunkt kunne der være en vis fordel ved at tilbyde indledende undervisning og kurser vedrørende (ikke-trykt) digital pligtaflevering til nybegyndere fra begge grupper.

## Vigtigheden af samarbejde

Sluttelig rummer undersøgelsen en række positive tilkendegivelser, som påpeger behovet for og værdien af samarbejde på tværs af praksisgrupper, og især hvordan et sådant samarbejde gavner begge grupper i forhold til at løse nogle af de ovenfor nævnte udfordringer. Vi må dog erkende, at webarkiveringsorganisationer og -institutioner måske ikke har ressourcer til at yde den nødvendige støtte til forskere. Der er en række forskellige årsager til dette, det kan bl.a. være "på grund af en blanding af kuratoriske, tekniske, juridiske, økonomiske og organisatoriske begrænsninger" (Brügger, 2021, s. 217). Sådanne faktorer kan være yderligere påvirket af de politiske og økonomiske omstændigheder i visse lande, som måske ikke er gunstige i forhold til finansieringen af kulturarvsprojekter, eller - på grund manglende arkiveringskapacitet - til at fremhæve arkivernes værdi over for interessenter (dvs. gennem caseundersøgelser på brugerniveau.) Dette udgør faktisk et paradoks, hvor webarkiveringsorganisationer har brug for ressourcer til at hjælpe forskere med at udvikle brugerundersøgelser for at demonstrere værdien af webarkiver til at opnå finansiering til at yde støtte til forskere. For organisationer, der ønsker at søge midler til at udvikle webarkiveringsinitiativer, er det således bydende nødvendigt at lave en forretningsmodel (fra starten) for aktiviteter i hele webarkiveringens livscyklus, herunder det at give adgang og at have støttemekanismer på plads til akademiske forskere eller andre slutbrugere som journalister eller advokater.

# RÉSUMÉ EXÉCUTIF

Cette étude fait partie d'un projet de collaboration entre des chercheurs de l'université de Maynooth, de la British Library, du Consortium international de préservation de l'Internet, de la Bibliothèque d'État de Bavière et de l'université de Siegen. Les membres de l'équipe de recherche sont tous membres du réseau d'études Web ARChive qui étudie les domaines et les événements du web (WARCnet, https://cc.au.dk/en/warcnet). WARCnet est financé par le Fonds de recherche indépendant Danemark | Humanités (subvention n° 9055-00005B).

Cette étude se concentre sur toute personne participant à la recherche des archives web, la curation et l'utilisation du contenu web archivé à des fins de recherche scientifique ou autres. En tant que telle, elle s'adresse à la fois aux créateurs et aux utilisateurs d'archives web. Nous estimons que la recherche sur les archives web est représentative des processus et activités décrits dans le modèle de cycle de vie de l'archivage Web d'Archive-It, depuis l'évaluation, l'acquisition et la préservation jusqu'à la relecture, l'accès, l'utilisation et la réutilisation (Bragg & Hannah, 2013). Cette étude a cherché à identifier et à documenter les compétences, les outils et les connaissances nécessaires pour atteindre un large éventail d'objectifs dans le cycle de vie de l'archivage web et à explorer les défis de la participation à la recherche sur les archives web ainsi que les interludes de ces défis à travers les communautés de pratique. Nous suggérons qu'il existe un besoin perpétuel d'examiner les rôles des compétences, des outils et des méthodes associés au cycle de vie de l'archivage web tant que les technologies internet, web et logiciels ne cessent de progresser, et d'évoluer.

La méthodologie de l'étude comprenait une recherche documentaire, la participation aux discussions des réunions du WARCnet et un questionnaire en ligne. Le questionnaire a été diffusé via les réseaux sociaux et par e-mail du 23 juillet au 21 septembre 2021. La stratégie de recrutement visait à cibler les archivistes, les bibliothécaires, les conservateurs, les gestionnaires de l'information, les universitaires, les chercheurs, les étudiants, les historiens, etc. et consistait en des messages sur les médias sociaux et des courriels de recrutement aux listes de réseaux pour les archivistes, les bibliothécaires, les conservateurs, les sciences humaines numériques, les etudes internet et les études sur les archives web. Les résultats se fondent sur un nombre final de 44 participants.

## Données démographiques

Dans cette étude, les participants (N=44) sont âgés de 18 à 64 ans et s'identifient comme résidant en Amérique du Nord, en Europe et en Asie. Les participants s'identifient comme

novices, intermédiaires et expérimentés dans le travail ou l'utilisation d'archives web, et il y a une représentation égale des participants qui s'identifient comme hommes et femmes. Cela peut indiquer que le sexe ne se présente pas comme un obstacle évident dans la recherche sur les archives web, du moins dans cette étude.

En ce qui concerne la position des participants, nous proposons deux représentations thématiques : (i) les participants qui ont déclaré travailler dans une bibliothèque, une archive ou un service d'archivage (du) web (n=30), et (ii) les participants qui ont déclaré être universitaires, maîtres de conférence, étudiants de troisième cycle ou de doctorat, ou travailler dans un cadre de développement informatique/web (n=14). Au départ, nous pensions qu'il serait possible d'aligner les positions des participants selon qu'ils étaient des créateurs d'archives web ou des utilisateurs d'archives web, mais ce ne fut pas le cas. En fait, les frontières étaient floues car certains répondants de la communauté de l'archivage web indiquent également qu'ils sont des utilisateurs d'archives web dans le cadre de leur travail. Alors que certains répondants de la communauté universitaire indiquent qu'ils sont des créateurs/curateurs d'archives web à des fins de recherche. Ainsi, la catégorisation des positions des participants n'était pas aussi tranchée qu'on l'imaginait à l'origine, et nous reconnaissons qu'il y a un certain chevauchement.

En se basant largement sur les intérêts, les antécédents, les expériences des participants et leurs relations avec la recherche sur les archives web, nous suggérons que les participants à cette étude s'identifient à un ou plusieurs des domaines suivants, par ordre alphabétique :

- Affaires et/ou droit
- Archives web, archivage web, curation
- Arts, sciences humaines, sciences humaines numériques, sciences sociales, études des médias
- Applications internet/web, systèmes
- Applications, systèmes, environnements informatiques/informatiques
- Science/analyse des données, statistiques
- Sciences de l'information (autres que l'archivage/curation du web).
- Utilisation des archives web et du contenu web archivé

## Principales conclusions et idées

Dans ce résumé, nous offrons une vue d'ensemble des conclusions et de la discussion, et l'organisons globalement en quatre sections principales comme suit :

- Compétences, connaissances, outils et méthodes dans la recherche d'archives web

- Défis liés à la recherche sur les archives web
- Les défis liés au dépôt légal, aux droits d'auteur et au GDPR
- Les collaborations sont essentielles

## Compétences, connaissances, outils et méthodes dans la recherche d'archives web.

À partir des résultats, nous avons présenté un large éventail de compétences, d'outils, de méthodes et de connaissances qui sont nécessaires, souhaitables ou utiles pour le domaine de la recherche sur les archives web à travers les communautés de pratique. Certaines des principales représentations comprennent :

- Logiciels et outils
- Archives web, archivage web, curation
- Programmation, langages de script
- Processus/flux de travail de la curation numérique
- Compétences en matière d'analyse de données
- Méthodes/approches de recherche
- Conception de sites web/compétences liées à internet
- Sciences de l'information (autres que l'archivage/curation du web).

Cette étude montre plusieurs points communs entre les participants qui se sont identifiés comme travaillant dans un environnement de bibliothèque, d'archives ou d'archives web, et les participants qui se sont identifiés comme étant érudits, universitaires, conférenciers, étudiants, ou travaillant dans un environnement de conception informatique/web. Par exemple, les répondants des deux communautés indiquent l'utilisation d'archives web pour trouver des informations, des documents et d'anciens sites web, et montrent des préoccupations similaires concernant les pertes et les changements dans le contenu web. Le traitement de volumes de données exceptionnellement importants est également mentionné comme un défi pour les répondants des deux communautés. Et les répondants des deux communautés indiquent l'importance d'acquérir des connaissances et des compétences techniques et critiques par le biais de formations, de cours et d'ateliers de travail, ainsi que par des collaborations et du mentorat. Ce qui apparaît également évident dans diverses sections des résultats, est le nombre de répondants des deux communautés qui indiquent le besoin de collaborations et de voies pour développer davantage de liens entre les créateurs/curateurs et les utilisateurs/chercheurs.

Sur le plan des outils et des méthodes, les deux communautés bénéficieraient d'une formation aux diverses méthodes de capture, notamment les logiciels d'exploration, les outils

de capture d'écran et de screencasting, ainsi que les outils de téléchargement de données à partir d'API. Il existe également des indications selon lesquelles le développement de matériel de formation à l'utilisation de logiciels de tableur, ainsi qu'à la gestion et à la préservation des tableurs en tant que sorties de données, serait utile pour les niveaux novice, intermédiaire et plus avancé dans l'ensemble de la communauté de recherche sur les archives Web. En outre, cette étude offre des indications selon lesquelles les utilisateurs d'archives Web bénéficieraient d'une formation d'introduction à l'archivage Web, tandis que le personnel dans un environnement d'archivage Web bénéficierait d'une certaine compréhension et d'une formation aux outils et méthodes utilisés par les utilisateurs/chercheurs pour analyser les données Web archivées. Nous devons toutefois souligner que cette étude montre que les participants issus d'un environnement érudit ou universitaire utilisent une diversité d'outils et de méthodes. De plus, la question ou la méthodologie de recherche influence souvent le choix des outils et des méthodes, par exemple, lorsque les données sont collectées manuellement pour une lecture attentive ou lorsque seules des parties spécifiques d'un site Web sont extraites. Ce groupe de participants est également confronté à des défis en raison du manque de méthodes de recherche, de théorie et d'approches pour combiner les méthodes traditionnelles avec la recherche sur les archives Web. Ainsi, les deux communautés bénéficieraient d'une formation commune collaborative en termes d'approches et de méthodes de recherche actuelles pour l'utilisation du Web archivé, incluant des démonstrations d'outils et de logiciels. De cette façon, le domaine serait enrichi grâce à l'apport du dialogue des deux communautés afin d'établir une meilleure compréhension des méthodes et approches de recherche pour l'utilisation des archives Web, ainsi que pour « acquérir une bonne compréhension du Web archivé en tant que type spécifique de source et des conséquences de ces caractéristiques » pour la recherche utilisant le web archivé, comme l'a souligné un répondant.

## Les défis que pose la recherche sur le web archivé

Cette étude identifie de multiples défis qui ont un impact sur l'ensemble des communautés de pratique. Par exemple, les défis liés à la saisie de contenu web dynamique entraînent souvent des lacunes en matière d'archivage, lacunes qui peuvent ensuite se manifester par des copies d'archives incohérentes et incomplètes pour l'utilisateur final. Les questions d'incomplétude dues à des actifs manquants ou à des liens brisés sur des sites web dynamiques sont problématiques à la fois pour les archivistes web et les utilisateurs finaux, en particulier lorsque les lacunes sont difficiles à documenter et à expliquer aux utilisateurs. La production de métadonnées et de documentation complètes pour les collections d'archives web représente un énorme défi pour les institutions d'archivage, car il s'agit d'un

processus qui prend beaucoup de temps et demande beaucoup de travail, exacerbé par l'énorme échelle des données. Des métadonnées et une documentation moins complètes sont ensuite problématiques pour l'utilisateur final qui cherche à s'engager avec les collections. En outre, un manque de ressources et de compétences spécialisées peut également affecter le développement d'une documentation complète, qui faciliterait la diversité des utilisateurs, qui ont en outre différents niveaux de compétences et d'expérience. Il faut également tenir compte du fait que les chercheurs universitaires et les autres utilisateurs finaux, tels que les journalistes ou les avocats, n'ont peut-être ni le temps, ni l'énergie d'investir dans l'acquisition d'une bonne compréhension de ces questions, ce qui peut être perçu comme une barrière à l'entrée ou un défi à l'engagement dans les archives web. Ainsi, il serait avantageux de fournir aux utilisateurs et aux utilisateurs potentiels une formation d'introduction à l'archivage web, dans un contexte localisé par rapport à l'archive web utilisée, dans le but d'offrir une plus grande sensibilisation, et donc une meilleure compréhension de l'étendue des collections par rapport aux limites des stratégies d'archivage dues aux défis techniques, aux contraintes légales et au manque de ressources. Il s'agit également d'une opportunité de collaboration entre les archives web et leurs utilisateurs pour développer une documentation à l'unisson, qui pourrait éventuellement être adaptée à toutes les disciplines et professions. Ceci serait un gain important pour les deux communautés, créant un cercle vertueux de création et d'utilisation finale.

Les répondants des deux communautés rencontrent des difficultés à acquérir de nouvelles compétences. Nous soulignons comment les deux communautés bénéficieraient de l'offre d'une formation commune collaborative sur l'ensemble des activités du cycle de vie de l'archivage web. Cette étude offre une vue d'ensemble des types de compétences et de connaissances que les praticiens et les utilisateurs d'archives web possédaient avant de travailler avec des archives web, des compétences qu'ils ont développées en travaillant avec des archives web et des défis auxquels ils ont été confrontés en travaillant avec ce type de ressource. Nous proposons que ces informations soient utilisées comme point de départ pour favoriser les discussions sur le développement de matériel de formation efficace pour les compétences et les outils nécessaires au travail avec les archives web, que ce soit en tant que créateur, conservateur, technicien ou utilisateur/chercheur. Nous suggérons en outre qu'une telle formation devra également être référencée dans une matrice de compétences, car il est très difficile de développer et de fournir une formation adéquate sans un point de référence auquel se mesurer. Nous constatons également que les défis rencontrés par les participants à cette étude ne s'atténuent pas avec l'augmentation de l'expérience et soulignent la nécessité d'une formation pour tous les niveaux d'expérience. Nous suggérons que, afin de développer des ressources ciblées pour les formations d'introduction et plus avancées, des

recherches supplémentaires seraient nécessaires pour voir comment les défis évoluent avec l'augmentation de l'expérience dans les communautés.

## Les défis liés au dépôt légal, aux droits d'auteur et au GDPR

Les défis liés aux aspects juridiques, tels que le dépôt légal, le droit d'auteur et le GDPR, constituent des obstacles pour les communautés d'archivage web et de chercheurs/utilisateurs. Les répondants des deux groupes discutent des difficultés à citer le contenu web archivé provenant d'archives de dépôt légal ou d'archives dont l'accès est restreint. Les participants qui se sont identifiés à la communauté de l'archivage web mentionnent des défis pour donner accès aux collections web archivées en raison de la législation, des droits d'auteur, du GDPR et des embargos. Les défis dus aux faibles taux de réponse au niveau de l'acquisition des permissions des propriétaires de sites web, sont également mentionnés, tant pour la capture des sites, que pour fournir l'accès aux sites archivés en dehors d'un bâtiment physique. On souligne également le fait que si le dépôt légal peut permettre la collecte de sites web par une institution de dépôt légal, il ne traite souvent pas efficacement de la fourniture d'accès. Pour certaines institutions, l'accès peut n'être fourni que sur place, ce qui « les rend économiquement inaccessible comme l'a noté un répondant. Il s'agit d'un domaine de recherche plus ciblé, car très peu d'attention a été accordée aux facteurs socio-économiques qui pourraient influencer les obstacles à l'entrée et à l'engagement dans les archives web.

Les participants qui s'identifient à la communauté académique discutent des défis liés à l'utilisation des archives web en raison des aspects juridiques en termes d'accès aux données, d'utilisation des données et de stockage des données provenant des archives web. D'autres défis incluent la manipulation de données protégées par le droit d'auteur à partir d'une archive web, ainsi que l'impossibilité de télécharger des données à partir de certaines archives web. Le travail sur des projets de collaboration transnationaux présente également des difficultés en raison des lois sur le dépôt légal qui varient d'un pays à l'autre et qui affectent la manière dont les données sont accessibles, utilisées et par qui. En outre, les difficultés à partager les données des archives web ou à les rendre réutilisables vont à l'encontre des tendances actuelles des bailleurs de fonds qui stipulent de plus en plus des cadres d'accès ouvert et de science ouverte pour les résultats de la recherche et des données. Nous suggérons que des discussions et une collaboration plus approfondies sont nécessaires, afin de favoriser le développement de l'application des pratiques de gestion des données de recherche dans les cadres de dépôt légal, les cadres scientifiques ouverts et les environnements de recherche d'archives web. Pour commencer, il serait utile de proposer

des formations et des cours d'introduction au dépôt légal numérique (non imprimé) aux novices des deux communautés.

## Les collaborations sont essentielles

Finalement, cette étude trouve des reconnaissances positives qui renforcent la nécessité et la valeur des collaborations entre les communautés de pratique, et surtout la façon dont ces collaborations profitent aux deux communautés pour relever certains des défis mentionnés ci-dessus. Cependant, nous devons reconnaître que les organisations et institutions d'archivage web peuvent ne pas avoir les ressources nécessaires pour fournir le soutien nécessaire aux chercheurs. Les raisons en sont variées et peuvent être « dues à un ensemble de contraintes curatoriales, techniques, juridiques, économiques et organisationnelles » (Brügger, 2021, p. 217). Ces facteurs peuvent être encore influencés par les climats politique et économique de certains pays qui peuvent ne pas être favorables au financement de projets de patrimoine culturel, ou en raison d'un manque de capacité de l'archivage web à promouvoir la valeur des archives web auprès des parties prenantes (c'est-à-dire par le biais d'études de cas d'utilisateurs). En effet, cela présente un paradoxe, à savoir que les organisations d'archivage web ont besoin de ressources pour aider les chercheurs à développer des études de cas d'utilisateurs afin de démontrer la valeur des archives web pour obtenir un financement afin de fournir un soutien aux chercheurs. Ainsi, pour les organisations qui souhaitent solliciter des fonds pour développer des initiatives d'archivage web, il est impératif de faire une analyse de rentabilité (dès le départ) pour les activités du cycle de vie complet de l'archivage web, y compris la fourniture de mécanismes d'accès et de soutien aux chercheurs universitaires, ou à d'autres utilisateurs finaux tels que les journalistes ou les avocats.

# RESUMEN

Este estudio forma parte de un proyecto colaborativo llevado a cabo por personal investigador de la Universidad de Maynooth, la Biblioteca Británica, el Consorcio Internacional para la Preservación de Internet, la Biblioteca Estatal de Baviera y la Universidad de Siegen. El equipo de investigación está integrado al completo por integrantes de la red de estudios WebARChive, dedicada a la investigación de dominios y eventos web (WARCnet, warcnet.eu). WARCnet está financiada por el Independent Research Fund Denmark | Humanities (subvención núm. 9055-00005B).

El estudio se centra en personas de todo el mundo que participan en la investigación de archivos web, en el contexto del archivado, la curación de contenidos y el uso de archivos web y contenido web archivado, con fines de investigación o de otra índole. Por consiguiente, va dirigido tanto a personas creadoras como a personas usuarias de archivos web. Consideramos que la investigación de archivos web es representativa de los procesos y actividades descritos en el modelo de ciclo de vida de archivado web de Archive-It, desde la valoración, adquisición y preservación hasta la reproducibilidad, el acceso, el uso y la reutilización (Bragg & Hannah, 2013). El propósito del estudio era identificar y documentar las habilidades, las herramientas y los conocimientos requeridos para alcanzar una amplia variedad de objetivos dentro del ciclo de vida del archivado web y explorar los retos que afectan a la participación en la investigación de archivos web, así como todo lo relacionado con dichos retos en todas las comunidades de práctica. Sugerimos que existe una necesidad constante de examinar los roles de las habilidades, las herramientas y los métodos asociados al ciclo de vida del archivado web mientras Internet y las tecnologías web y de software sigan avanzando, mejorando y cambiando.

La metodología para el estudio implicó investigación sobre datos secundarios (*desk research*), la participación en debates en reuniones de WARCnet y la realización de un cuestionario online. El cuestionario se difundió a través de redes sociales y mediante correo electrónico entre el 23 de julio y el 21 de septiembre de 2021. La estrategia de captación iba dirigida a archivistas, personal bibliotecario, personas curadoras de contenidos, personal gestor de información, personal académico, personal investigador, estudiantado, personal historiador, etc. y consistía en publicaciones en redes sociales y la captación de correos electrónicos en listas de distribución de archivistas, personal bibliotecario, personas curadoras de contenidos, humanidades digitales, estudios de Internet y estudios de archivado web. Los resultados se basan en un recuento final de 44 participantes.

# Datos demográficos

Las personas participantes (N=44) en este estudio tienen una edad comprendida entre los 18 y los 64 años y se identifican como residentes en América del Norte, Europa y Asia. Las personas participantes se identifican con niveles de principiante, intermedio y experto en el trabajo o el uso de archivos web, y existe una representación equitativa entre participantes que se identifican como hombres y como mujeres. Esto podría indicar que el género no se presenta como una barrera obvia en la investigación de archivos web, o al menos no en este estudio.

Por lo que se refiere al puesto de trabajo de las personas participantes, ofrecemos dos representaciones temáticas: (i) participantes que se identificaron con el trabajo en una biblioteca, archivo o entorno de archivo web (n=30) y (ii) participantes que se identificaron como intelectuales, personal académico, profesorado universitario, estudiantes de posgrado o doctorado o personas cuyo trabajo está relacionado con las tecnologías de la información/un entorno relacionado con el diseño web (n=14). En un principio pensamos que sería posible alinear los puestos de trabajo de las personas participantes en función de si estos eran personas creadoras o personas usuarias de archivos web, pero no fue posible. De hecho, los límites no estaban claros, puesto que algunos de las personas encuestadas de la comunidad de archivado web aseguran ser al mismo tiempo usuarias de archivos web como parte de su trabajo, mientras que algunas personas encuestadas de la comunidad académica se identifican como personal creador/personal *curador* de archivos web a efectos de investigación. Así pues, la categorización de los puestos de trabajo de las personas participantes no estaba tan definida como imaginamos en un principio, y reconocemos que existe cierto solapamiento.

Basándonos ampliamente en los intereses, el historial y las experiencias de las personas participantes, así como en su relación con la investigación de archivado web, sugerimos que las personas participantes de este estudio se identifiquen con una o varias de las siguientes áreas temáticas, por orden alfabético.

- Archivos web, archivado de webs, *curación*
- Artes, Ciencias sociales, Estudios de medios de comunicación, Humanidades, Humanidades digitales
- Ciencia/análisis de datos, Estadística
- Ciencias de la información (distintas del archivado/*curación* de webs)
- Internet/aplicaciones web, sistemas
- Negocios o Derecho
- TI/Aplicaciones, sistemas y entornos informáticos

- Uso de archivos web y contenido web archivado

# Principales descubrimientos y conclusiones

En el presente resumen ofrecemos una visión general de las conclusiones y el debate y los organizamos en líneas generales en cuatro secciones principales, a saber:

- Habilidades, conocimientos, herramientas y métodos en la investigación de archivos web
- Retos relacionados con la investigación de archivos web
- Retos relacionados con el depósito legal, el *copyright* y el RGPD
- Las colaboraciones son clave

## Habilidades, conocimientos, herramientas y métodos en la investigación de archivos web

A partir de las conclusiones, presentamos una amplia variedad de habilidades, herramientas, métodos y conocimientos necesarios, deseables o útiles en el área de la investigación de archivos web en todas las comunidades de práctica. Algunas de las representaciones principales son:

- Software y herramientas
- Archivos web, archivado de webs, *curación*
- Programación, lenguajes de programación
- Procesos de *curación* digital/flujos de trabajo
- Habilidades de análisis de datos
- Métodos/enfoques de investigación
- Diseño web/habilidades relacionadas con Internet
- Ciencias de la información (distintas del archivado/*curación* de webs)

El estudio revela varios puntos en común entre las personas participantes que se identificaron con el trabajo en una biblioteca, archivo o entorno de archivos web, y las personas participantes que se identificaron como personal académico, intelectuales, profesorado universitario, estudiantado o personas que trabajan en un entorno de TI/diseño web. Por ejemplo, las personas encuestadas de ambas comunidades mencionan el uso de archivos web para buscar información, bibliografía y sitios web antiguos, y muestran una preocupación similar por las pérdidas y los cambios en el contenido web. Otro de los retos que mencionan las personas encuestadas de ambas comunidades es el de la gestión de volúmenes de datos excepcionalmente grandes. Y las personas encuestadas de ambas comunidades indican la importancia de adquirir conocimientos y habilidades técnicas y críticas mediante formación,

cursos y talleres, así como a través de colaboraciones y mentorías. Otro aspecto que también parece evidente a partir de varias secciones de los resultados es el número de personas encuestadas de ambas comunidades que señalan la necesidad de colaboraciones y vías para establecer nuevas conexiones entre personas creadoras/personas curadoras y personas usuarias/personal investigador.

En lo que se refiere a herramientas y métodos, ambas comunidades se beneficiarían de una formación en diversos métodos de captura, como software de rastreo o *crawling*, herramientas para hacer pantallazos y capturas de pantalla y videografías (*screencasting*), así como herramientas para la descarga de datos de API. También se señala que el desarrollo de materiales de formación en el uso de software de hojas de cálculo y la gestión y preservación de hojas de cálculo como salida de datos sería útil para las personas participantes de nivel principiante, intermedio y más avanzado en toda la comunidad de investigación de archivos web en su conjunto. Asimismo, el estudio apunta a que las personas usuarias de archivos web se beneficiarían de una formación de introducción al archivado web, mientras que al personal que trabaja en un entorno de archivado web le beneficiaría conocer y formarse en las herramientas y los métodos que utilizan las personas usuarias/personal investigador para analizar los datos de web archivadas. No obstante, cabe señalar que según el estudio las personas participantes que proceden de un entorno universitario o académico utilizan variedad de herramientas y métodos. Además, la pregunta o la metodología de investigación suele influir en la elección de herramientas y métodos; por ejemplo, si los datos se recogen manualmente para una lectura exhaustiva o cuando solo se extrae información de determinadas secciones de un sitio web (*web scraping*). Este grupo de participantes suele enfrentarse también a algunos retos derivados de la falta de métodos, teoría y enfoques de investigación para combinar los métodos tradicionales con la investigación de archivos web. Por tanto, ambas comunidades se beneficiarían de una formación común colaborativa sobre los enfoques y métodos actuales de investigación para el uso de la web archivada, incluidas demostraciones de herramientas y de software. De este modo, este campo se enriquecería con la aportación del diálogo entre ambas comunidades para desarrollar un mejor conocimiento de los métodos y enfoques de investigación para el uso de archivos web, además de «entender mejor las web archivadas como un tipo específico de fuente y las consecuencias de estas características para la investigación usando la web archivada», como comentó uno de las personas encuestadas.

## Retos relacionados con la investigación de archivos web

El estudio identifica múltiples retos que afectan a todas las comunidades de práctica. Por ejemplo, los retos relacionados con la captura de contenido web dinámico suelen derivar en deficiencias archivísticas que pueden traducirse en copias de archivo inconsistentes e incompletas para el usuario final. Las cuestiones relacionadas con archivos incompletos debido a activos que faltan o a enlaces rotos en sitios web *en vivo* suponen un problema tanto para personal archivista web como para las personas usuarias finales, especialmente cuando dicha falta de datos resulta difícil de documentar y explicar a las personas usuarias. La producción de metadatos y documentación completos para colecciones de archivos web presenta un enorme reto para las instituciones archivísticas, puesto que se trata de un proceso muy laborioso y que requiere mucha mano de obra, a lo que se suma la gigantesca escala de los datos. Unos metadatos y una documentación incompletos resultan problemáticos para las personas usuarias finales que buscan trabajar con las colecciones. Por otra parte, la falta de recursos y de especialización también podría afectar a la elaboración de una documentación completa, que facilitaría la diversidad de personas usuarias con distintos niveles de habilidades y de experiencia. También hay que tener en cuenta que el personal investigador académico y otras personas usuarias finales (como periodistas o juristas) podrían no disponer del tiempo o la energía que necesitarían invertir para comprender bien estos problemas y, por consiguiente, esto podría percibirse como una barrera a los archivos web o un reto a la hora de interactuar con ellos. Así pues, sería ventajoso ofrecer a las personas usuarias y posibles personas usuarias una formación introductoria sobre archivado web, en un contexto localizado, en relación con el archivo web que están utilizando, en un intento por fomentar la concienciación y, por ende, una mejor comprensión del alcance de las colecciones frente a las limitaciones de las estrategias archivísticas que imponen los retos técnicos, las restricciones legales y la falta de recursos. También brinda una oportunidad de colaboración entre archivos web y sus personas usuarias para desarrollar una documentación de manera simultánea, que en última instancia podría adaptarse a las distintas disciplinas y profesiones. Esto sería una ventaja significativa para ambas comunidades, y establecería un círculo «virtuoso» de creación y uso final.

Las personas encuestadas de ambas comunidades experimentan dificultades en el aprendizaje de nuevas destrezas. Queremos destacar las ventajas que tendría para ambas comunidades la impartición de una formación comunitaria colaborativa en toda la serie de actividades incluidas en el ciclo de vida del archivado web. El estudio ofrece una visión general de los tipos de habilidades y conocimientos que tenían profesionales del archivado web y personas usuarias de archivos web antes de trabajar en archivos web; las habilidades que

adquirieron mientras trabajaban con archivos web y los retos a los que se enfrentaron al trabajar con este tipo de recursos. Proponemos que esto sirva de punto de partida para fomentar el debate sobre el desarrollo de materiales formativos efectivos para adquirir las habilidades y herramientas necesarias para trabajar con archivos web en todo el espectro de personas creadoras, personas *curadoras*, personal técnico y personas usuarias/personal investigador. Asimismo, sugerimos que dicha formación deberá medirse en una matriz de habilidades, puesto que es muy difícil desarrollar e impartir la formación adecuada sin una referencia frente a la que se pueda medir. También concluimos que los retos que experimentan quienes participaron en el estudio no se reducen a medida que aumenta su experiencia, y destacamos la necesidad de formación en todos los niveles de experiencia. Sugerimos que, para desarrollar unos recursos bien enfocados, ya sea para una formación introductoria o más avanzada, se requeriría una mayor investigación para observar cómo cambian los retos a medida que aumenta la experiencia en todas las comunidades.

## Retos relacionados con el depósito legal, el *copyright* y el RGPD

Los retos relativos a aspectos legales, como el depósito legal, el *copyright* y el RGPD presentan obstáculos tanto para el archivado de webs como para las comunidades de personal investigador/ personas usuarias. Las personas encuestadas de los dos grupos hablan sobre los retos para citar contenido web archivado procedente de archivos de un depósito legal o de archivos de acceso restringido. Las personas participantes que se identificaron con la comunidad de archivado web mencionan los retos que limitan el acceso a colecciones de webs archivadas debido a la legislación, el *copyright*, el RGPD y los embargos. También se mencionan los retos derivados de los bajos índices de respuesta en la obtención de permisos concedidos por propietarios de sitios web, tanto para la captura de sitios como en la concesión de acceso a los sitios archivados fuera de un edificio físico. Además, destaca el hecho de que si bien el depósito legal podría permitir la recogida de sitios web por parte de una institución de depósito legal, a menudo no trata de manera eficaz la cuestión del acceso. Para algunas instituciones, el acceso solo es posible *in situ*, «lo que las convierte en económicamente inaccesibles», en palabras de una de las personas participantes en la encuesta. Esta área requiere una investigación mejor enfocada, puesto que se ha prestado muy poca atención a los factores socio-económicos que podrían suponer unas barreras frente a los archivos web y frenar la interacción con los mismos.

Las personas participantes que se identificaron con la comunidad académica hablan sobre los retos de usar archivos web debidos a cuestiones legales en cuanto a acceso, uso y almacenamiento de los datos procedentes de archivos web. Otros retos son la gestión de

datos protegidos por *copyright* procedentes de un archivo web, así como la incapacidad de descargar datos de determinados archivos web. También se identifican retos relacionados con el trabajo en proyectos colaborativos transnacionales, debido a las diferencias en las leyes de depósito legal entre países que afectan al modo de usar y acceder a los datos y a quién puede hacerlo. Además, los retos relacionados con el intercambio de datos procedentes de archivos web o con su reutilización son contrarios a las tendencias actuales de las entidades financiadoras, que cada vez más estipulan el acceso abierto y los llamados O*pen Science Frameworks* (marcos de ciencia abierta) para la investigación y la generación de datos. Consideramos que es necesario un mayor debate y una mayor colaboración para fomentar el desarrollo en la aplicación de las prácticas de gestión de datos de investigación dentro de los marcos del depósito legal, los *Open Science Frameworks* y los entornos de investigación de archivos web. Como punto de partida, sería beneficioso ofrecer formación y cursos introductorios relacionados con el depósito legal digital (no impreso) dirigido a principiantes de ambas comunidades.

## Las colaboraciones son clave

Finalmente, el estudio encuentra reconocimientos positivos que refuerzan la necesidad y el valor de las colaboraciones entre comunidades de práctica, y especialmente cómo tales colaboraciones benefician a ambas comunidades al abordar algunos de los retos previamente mencionados. No obstante, debemos reconocer que las organizaciones e instituciones dedicadas al archivado de webs podrían no disponer de los recursos para ofrecer el apoyo necesario al personal investigador. Los motivos son varios y pueden «deberse a una combinación de restricciones de *curaduría*, restricciones técnicas, jurídicas, económicas y organizativas». (Brügger, 2021, p. 217). Dichos factores podría verse afectados también por el clima político y económico de determinados países, lo cual puede ser desfavorable para la financiación de proyectos de patrimonio cultural, o bien deberse a una falta de capacidad de archivado web para promover el valor de los archivos web a los agentes implicados (es decir, a través de estudios de caso de usuarios). De hecho, esto presenta una paradoja, por cuanto las organizaciones dedicadas al archivado web necesitan recursos para ayudar al personal investigador a desarrollar estudios de caso de personas usuarias que les permitan demostrar el valor de los archivos web para conseguir financiación que apoye al personal investigador. Así pues, para las organizaciones que desean buscar financiación para desarrollar iniciativas de archivado web es esencial crear un caso de negocio (desde el inicio) para actividades en el ciclo de vida completo de archivado web, incluso ofrecer mecanismos de acceso y soporte para personal investigador académico u otras personas usuarias finales, como periodistas o juristas.

# RESUM

Aquest estudi forma part d'un projecte col·laboratiu dut a terme per personal investigador de la Universitat de Maynooth, la Biblioteca Britànica, el Consorci Internacional per a la Preservació d'Internet, la Biblioteca Estatal de Baviera i la Universitat de Siegen. L'equip de recerca està integrat al complet per integrants de la xarxa d'estudis WebARChive, dedicada a la investigació de dominis i esdeveniments web (WARCnet, warcnet.eu). WARCnet està finançada per l'Independent Research Fund Denmark | Humanities (subvenció núm. 9055-00005B).

L'estudi se centra en persones de tot el món que participen en la recerca d'arxius web, en el context de l'arxivat, curació de continguts i l'ús d'arxius web i contingut web arxivat, amb finalitats de recerca o d'una altra índole. Per tant, va dirigit tant a persones creadores com a persones usuàries d'arxius web. Considerem que la recerca d'arxius web és representativa dels processos i activitats descrits en el model de cicle de vida d'arxivat web d'Archive-It, des de la valoració, adquisició i preservació fins a la reproductibilitat, l'accés, l'ús i la reutilització (Bragg & Hannah, 2013). El propòsit de l' estudi era identificar i documentar les habilitats, les eines i els coneixements requerits per assolir una àmplia varietat d' objectius dins del cicle de vida de l' arxivat web i explorar els reptes que afecten la participació en la recerca d' arxius web, així com tot allò relacionat amb aquests reptes en totes les comunitats de pràctica. Suggerim que existeix una necessitat constant d'examinar els rols de les habilitats, les eines i els mètodes associats al cicle de vida de l'arxivat web mentre Internet i les tecnologies web i de programari continuïn avançant, millorant i canviant.

La metodologia per a l'estudi va *implicar* recerca sobre dades secundàries (*desk resea* r *ch*), la participació en debats en reunions de WARCnet i la realització d'un qüestionari online. El qüestionari es va difondre a través de xarxes socials i per correu electrònic entre el 23 de juliol i el 21 de setembre de 2021. L'estratègia de captació anava dirigida a arxivistes, personal bibliotecari, persones curadores de continguts, personal gestor d'informació, personal acadèmic, personal investigador, estudiantat, personal historiador, etc. i consistia en publicacions en xarxes socials i la captació de correus electrònics en llistes de distribució d'arxivistes, personal bibliotecari, persones curadores de continguts, humanitats digitals, estudis d'Internet i estudis d'arxivat web. Els resultats es basen en un recompte final de 44 participants.

# Dades demogràfiques

Les persones participants (N=44) en aquest estudi tenen una edat compresa entre els 18 i els 64 anys i s'identifiquen com a residents a Amèrica del Nord, Europa i Àsia. Les persones participants s' identifiquen amb nivells de principiant, intermedi i expert en el treball o l' ús d' arxius web, i existeix una representació equitativa entre participants que s' identifiquen com a homes i com a dones. Això podria indicar que el gènere no es presenta com una barrera òbvia en la investigació d'arxius web, o almenys no en aquest estudi.

Pel que fa al lloc de treball de les persones participants, oferim dues representacions temàtiques: (i) participants que es van identificar amb el treball en una biblioteca, arxiu o entorn d'arxiu web (n=30) i (ii) participants que es van identificar com a intel·lectuals, personal acadèmic, professorat universitari , estudiants de postgrau o doctorat o persones el treball de les quals està relacionat amb les tecnologies de la informació / un entorn relacionat amb el disseny web (n=14). En un principi pensem que seria possible alinear els llocs de treball de les persones participants en funció de si aquests eren persones usuàries d'arxius web, però no va ser possible. De fet, els límits no estaven clars, ja que alguns de les persones enquestades de la comunitat d'arxivat web asseguren ser alhora usuàries d'arxius web com a part del seu treball, mentre que algunes persones enquestades de la comunitat acadèmica s'identifiquen com a personal creador/personal *curador* d' arxius web a efectes de recerca. Així doncs, la categorització dels llocs de treball de les persones participants no estava tan definida com imaginem en un principi, i reconeixem que hi ha cert solapament.

Basant-nos àmpliament en els interessos, l' historial i les experiències de les persones participants, així com en la seva relació amb la recerca d' arxivat web, suggerim que les persones participants d' aquest estudi s' identifiquin amb una o diverses de les següents àrees temàtiques, per ordre alfabètic.

- Arts, Ciències socials, Estudis de mitjans de comunicació, Humanitats, Humanitats digitals
- Arxius web, arxivat de webs, *curació*
- Ciència/anàlisi de dades, Estadística
- Ciències de la informació (diferents de l'arxivat/*curació* de webs)
- Internet/aplicacions web, sistemes
- Negocis o Dret
- TI/Aplicacions, sistemes i entorns informàtics
- Ús d'arxius web i contingut web arxivat

# Principals descobriments i conclusions

En el present resum oferim una visió general de les conclusions i el debat i els organitzem en línies generals en quatre seccions principals, a saber:

- Habilitats, coneixements, eines i mètodes en la recerca d' arxius web
- Reptes relacionats amb la recerca d' arxius web
- Reptes relacionats amb el dipòsit legal, el *copyright* i el RGPD
- Les col·laboracions són clau

## Habilitats, coneixements, eines i mètodes en la recerca d' arxius web

A partir de les conclusions, presentem una àmplia varietat d' habilitats, eines, mètodes i coneixements necessaris, desitjables o útils en l' àrea de la recerca d' arxius web en totes les comunitats de pràctica. Algunes de les representacions principals són:

- Programari i eines
- Arxius web, arxivat de webs, *curació*
- Programació, llenguatges de programació
- Processos de *curació* digital/fluxos de treball
- Habilitats d' anàlisi de dades
- Mètodes/enfocaments de recerca
- Disseny web/habilitats relacionades amb Internet
- Ciències de la informació (diferents de l'arxivat/*curació* de webs)

L'estudi revela diversos punts en comú entre les persones participants que es van identificar amb el treball en una biblioteca, arxiu o entorn d'arxius web, i les persones participants que es van identificar com a personal acadèmic, intel·lectuals, professorat universitari, estudiants o persones que treballen en un entorn de TI/disseny web. Per exemple, les persones enquestades d'ambdues comunitats esmenten l'ús d'arxius web per buscar informació, bibliografia i llocs web antics, i mostren una preocupació similar per les pèrdues i els canvis en el contingut web. Un altre dels reptes que esmenten les persones enquestades d'ambdues comunitats és el de la gestió de volums de dades excepcionalment grans. I les persones enquestades d'ambdues comunitats indiquen la importància d'adquirir coneixements i habilitats tècniques i crítiques mitjançant formació, cursos i tallers, així com a través de col·laboracions i mentories. Un altre aspecte que també sembla evident a partir de diverses seccions dels resultats és el nombre de persones enquestades d'ambdues comunitats que assenyalen la necessitat de col·laboracions i vies per establir noves connexions entre persones creadores/persones curadores i persones usuàries/personal investigador.

Pel que fa a eines i mètodes, ambdues comunitats es beneficiarien d'una formació en diversos mètodes de captura, com programari de rastreig o *crawling*, eines per fer pantalles i captures de pantalla i videografies (*screencasting*), així com eines per a la descàrrega de dades d'API. També s' assenyala que el desenvolupament de materials de formació en l' ús de programari de fulls de càlcul i la gestió i preservació de fulls de càlcul com a sortida de dades seria útil per a les persones participants de nivell principiant, intermedi i més avançat en tota la comunitat de recerca d' arxius web en el seu conjunt. Així mateix, l'estudi apunta que les persones usuàries d'arxius web es beneficiarien d'una formació d'introducció a l'arxivat web, mentre que al personal que treballa en un entorn de web li beneficiaria conèixer i formar-se en les eines i els mètodes que utilitzen les persones usuàries/personal investigador per analitzar les dades de web arxivades. No obstant això, cal assenyalar que segons l' estudi les persones participants que procedeixen d' un entorn universitari o acadèmic utilitzen varietat d' eines i mètodes. A més, la pregunta o la metodologia d'investigació sol influir en l'elecció d'eines i mètodes; per exemple, si les dades es recullen manualment per a una lectura exhaustiva o quan només s'extreu informació de determinades seccions d'un lloc web (*web scraping*). Aquest grup de participants sol enfrontar-se també a alguns reptes derivats de la falta de mètodes, teoria i enfocaments de recerca per combinar els mètodes tradicionals amb la recerca d'arxius web. Per tant, ambdues comunitats es beneficiarien d'una formació comuna col·laborativa sobre els enfocaments i mètodes actuals de recerca per a l'ús de la web arxivada, incloses demostracions d'eines i de programari. D'aquesta manera, aquest camp s'enriquiria amb l'aportació del diàleg entre ambdues comunitats per desenvolupar un millor coneixement dels mètodes i enfocaments d'investigació per a l'ús d'arxius web, a més d'«entendre millor les web arxivades com un tipus específic de font i les conseqüències d'aquestes característiques per a la recerca usant la web arxivada», com va comentar un de les persones enquestades.

## Reptes relacionats amb la recerca d' arxius web

L' estudi identifica múltiples reptes que afecten totes les comunitats de pràctica. Per exemple, els reptes relacionats amb la captura de contingut web dinàmic solen derivar en deficiències arxivístiques que poden traduir-se en còpies d'arxiu inconsistents i incompletes per a l'usuari final. Les qüestions relacionades amb arxius incomplets a causa d'actius que falten o a enllaços trencats en llocs web *en viu* suposen un problema tant per a personal web com per a les persones usuàries finals, especialment quan aquesta falta de dades resulta difícil de documentar i explicar a les persones usuàries . La producció de metadades i documentació completes per a col·leccions d'arxius web presenta un enorme repte per a les institucions arxivístiques, ja que es tracta d'un procés molt laboriós i que requereix molta mà d'obra, a la

qual cosa se suma la gegantina escala de les dades. Unes metadades i una documentació incompletes resulten problemàtics per a les persones usuàries finals que busquen treballar amb les col·leccions. D' altra banda, la manca de recursos i d' especialització també podria afectar l' elaboració d' una documentació completa, que facilitaria la diversitat de persones usuàries amb diferents nivells d' habilitats i d' experiència. També cal tenir en compte que el personal investigador acadèmic i otrapersones usuàries finals (com periodistes o juristes) podrien no disposar del temps o l'energia que necessitarien invertir per comprendre bé aquests problemes i, per tant, això podria percebre's com una barrera als arxius web o un repte a l'hora d'interactuar amb ells. Així doncs, seria avantatjós oferir a les persones usuàries i possibles persones usuàries una formació introductòria sobre el web, en un context localitzat, en relació amb l'arxiu web que estan utilitzant, en un intent per fomentar la conscienciació i, per tant, una millor comprensió de l'abast de les col·leccions enfront de les limitacions de les estratègies arxivístiques que imposen els reptes tècnics, les restriccions legals i la manca de recursos. També brinda una oportunitat de col·laboració entre arxius web i les seves persones usuàries per desenvolupar una documentació de manera simultània, que en última instància podria adaptar-se a les diferents disciplines i professions. Això seria un avantatge significatiu per a totes dues comunitats, i establiria un cercle «virtuós» de creació i ús final.

Les persones enquestades d' ambdues comunitats experimenten dificultats en l' aprenentatge de noves destreses. Volem destacar els avantatges que tindria per a ambdues comunitats la impartició d'una formació comunitària col·laborativa en tota la sèrie d'activitats incloses en el cicle de vida de l'arxivat web. L'estudi ofereix una visió general dels tipus d'habilitats i coneixements que tenien professionals del web i persones usuàries d'arxius web abans de treballar en arxius web; les habilitats que van adquirir mentre treballaven amb arxius web i els reptes als quals es van enfrontar en treballar amb aquest tipus de recursos. Proposem que això serveixi de punt de partida per fomentar el debat sobre el desenvolupament de materials formatius efectius per adquirir les habilitats i eines necessàries per treballar amb arxius web en tot l'espectre de persones creadores, persones *curadores*, personal tècnic i persones usuàries/personal investigador. Així mateix, suggerim que aquesta formació s' haurà de mesurar en una matriu d' habilitats, ja que és molt difícil desenvolupar i impartir la formació adequada sense una referència davant la qual es pugui mesurar. També concloem que els reptes que experimenten els qui van participar en l'estudi no es redueixen a mesura que augmenta la seva experiència, i destaquem la necessitat de formació en tots els nivells d'experiència. Suggerim que, per desenvolupar uns recursos ben enfocats, ja sigui per a una formació introductòria o més avançada, es requeriria una major

recerca per observar com canvien els reptes a mesura que augmenta l'experiència en totes les comunitats.

## Reptes relacionats amb el dipòsit legal, el *copyright* i el RGPD

Els reptes relatius a aspectes legals, com el dipòsit legal, el *copyright* i el RGPD presenten obstacles tant per a l' arxivat de webs com per a les comunitats de personal investigador/ persones usuàries. Les persones enquestades dels dos grups parlen sobre els reptes per citar contingut web arxivat procedent d'arxius d'un dipòsit legal o d'arxius d'accés restringit. Les persones participants que es van identificar amb la comunitat d'arxivat web esmenten els reptes que limiten l'accés a col·leccions de webs arxivades a causa de la legislació, el *copyright*, el RGPD i els embargaments. També s' esmenten els reptes derivats dels baixos índexs de resposta en l' obtenció de permisos concedits per propietaris de llocs web, tant per a la captura de llocs com en la concessió d' accés als llocs arxivats fora d' un edifici físic. A més, destaca el fet que si bé el dipòsit legal podria permetre la recollida de llocs web per part d' una institució de dipòsit legal, sovint no tracta de manera eficaç la qüestió de l' accés. Per a algunes institucions, l'accés només és possible *in situ*, «cosa que les converteix en econòmicament inaccessibles», en paraules d'una de les persones participants en l'enquesta. Aquesta àrea requereix una investigació més ben enfocada, ja que s' ha prestat molt poca atenció als factors socioeconòmics que podrien suposar unes barreres enfront dels arxius web i frenar-ne la interacció.

Les persones participants que es van identificar amb la comunitat acadèmica parlen sobre els reptes d'usar arxius web deguts a qüestions legals quant a accés, ús i emmagatzematge de les dades procedents d'arxius web. Altres reptes són la gestió de dades protegides per *copyright* procedents d'un arxiu web, així com la incapacitat de descarregar dades de determinats arxius web. També s'identifiquen reptes relacionats amb el treball en projectes col·laboratius transnacionals, a causa de les diferències en les lleis de dipòsit legal entre països que afecten la manera d'usar i accedir a les dades i a qui pot fer-ho. A més, els reptes relacionats amb l'intercanvi de dades procedents d'arxius web o amb la seva reutilització són contraris a les tendències actuals de les entitats finançadores, que cada vegada més estipulen l'accés obert i els anomenats O*pen Science Frameworks* (marcs de ciència oberta) per a la recerca i la generació de dades. Considerem que és necessari un major debat i una major col·laboració per fomentar el desenvolupament en l'aplicació de les pràctiques de gestió de dades de recerca dins dels marcs del dipòsit legal, els *Open Science Frameworks* i els entorns de recerca d'arxius web. Com a punt de partida, seria beneficiós oferir formació i cursos

introductoris relacionats amb el dipòsit legal digital (no imprès) dirigit a principiants d'ambdues comunitats.

## Les col·laboracions són clau

Finalment, l'estudi troba reconeixements positius que reforcen la necessitat i el valor de les col·laboracions entre comunitats de pràctica, i especialment com aquestes col·laboracions beneficien ambdues comunitats en abordar alguns dels reptes prèviament esmentats. No obstant això, hem de reconèixer que les organitzacions i institucions dedicades a l'arxivat de webs podrien no disposar dels recursos per oferir el suport necessari al personal investigador. Els motius són diversos i es poden «deure a una combinació de restriccions de *curadoria*, restriccions tècniques, jurídiques, econòmiques i organitzatives». (Brügger, 2021, p. 217). Aquests factors es podria veure afectats també pel clima polític i econòmic de determinats països, la qual cosa pot ser desfavorable per al finançament de projectes de patrimoni cultural, o bé deure's a una manca de capacitat d'arxivat web per promoure el valor dels arxius web als agents implicats (és a dir, a través d'estudis de cas d'usuaris). De fet, això presenta una paradoxa, ja que les organitzacions dedicades a l'arxivat web necessiten recursos per ajudar  el personal investigador a desenvolupar estudis de cas de persones usuàries que els permetin demostrar el valor dels arxius web per aconseguir finançament que doni suport al personal investigador . Així doncs, per a les organitzacions que volen buscar finançament per desenvolupar iniciatives d'arxivat web és essencial crear un cas de negoci (des de l'inici) per a activitats en el cicle de vida complet de web, fins i tot oferir mecanismes d'accés i suport per a personal investigador acadèmic o altres persones usuàries finals, com periodistes o juristes.

WARCnet
web archive studies

**WARCnet** *Special Reports* is a series of reports related to the activities of the WARCnet network. To ensure the relevance of the publications, WARCnet strives to publish with a rapid turnover. WARCnet Special Reports are edited by Niels Brügger, Jane Winters, Valérie Schafer, Kees Teszelszky, Peter Webster and Michael Kurzmeier. In cases where a WARCnet Special Report has gone through a process of single blind review, this is mentioned in the individual publication.

The aim of the WARCnet network is to promote high-quality national and transnational research that will help us to understand the history of (trans)national web domains and of transnational events on the web, drawing on the increasingly important digital cultural heritage held in national web archives. The network activities run in 2020-23, hosted by the School of Communication and Culture at Aarhus University, and are funded by the Independent Research Fund Denmark | Humanities (grant no 9055-00005B).