The Future(s) of Web Archive Research Across Ireland

Sharon C. Healy

Department of Computer Science

Faculty of Science and Engineering

National University of Ireland, Maynooth

Supervisors: Dr. Joseph Timoney

Dr. Martin Maguire

Dr. Juan-José Boté-Vericad

Mr. Thomas Lysaght

Head of Department: Dr. Joseph Timoney

A thesis submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy

October 2022

Funded by:

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Sharon C. Healy

October 2022

LIST OF CONTENTS

List of Figures

List of Tables

List of Abbreviations

AADDA	Analytical Access to the Domain Dark Archive
ADHO	Alliance of Digital Humanities Organizations
AOIR	Association of Internet Researchers
API	Application Programming Interface
ARC_IA	Internet Archive ARC file format
BnF	Bibliothèque nationale de France
BUDDAH	Big UK Domain Data for the Arts and Humanities project
BBS	bulletin board system
ccTLD	country code Top Level Domain
CD-ROM	Compact Disc-Read Only Memory
COIPLPA	Copyright and Other Intellectual Property Law Provisions Act 2019 [Ireland]
CRC	Copyright Review Committee [Ireland]
CRRA	Copyright and Related Rights Act, 2000 [Ireland]
DCC	Digital Curation Centre
DH	Digital Humanities
DIGLIB	Digital Libraries Research Mailing List
DMP	Data Management Plans
DOI	Digital Object Identifier
DVD	Digital Video Disc
EU	European Union
EC	European Communities
EEC	European Economic Community
GDPR	General Data Protection Regulation
HAW	Croatian Web Archive
HTML	HyperText Markup Language

IIPC	International Internet Preservation Consortium
IFLA	International Federation of Library Associations
ICPPA	Industrial and Commercial Property (Protection) Act, 1927 [Ireland]
IMF	Internet Memory Foundation
INA	Institut Nationale de l'Audiovisuel [France]
IRC	Irish Research Council
ISBN	International Standard Book Number
ISO	International Standardization Organisation
IT	Information Technology
JISC	Joint Information Systems Committee [United Kingdom]
KB	Koninklijke Bibliotheek [The Netherlands]
NAI	National Archives of Ireland
NI	Northern Ireland
NLI	National Library of Ireland
NDSA	National Digital Stewardship Alliance [United States]
NPLD	The Legal Deposit Libraries (Non-Print Works) Regulations 2013 [United Kingdom]
ODU WS-DL	Old Dominion University, Web Science and Digital Libraries Research Group [United States]
PANDORA	Preserving and Accessing Networked Documentary Resources of Australia
PDF	Portable Document Format
PNG	Portable Network Graphics
PROI	Public Record Office of Ireland
PRONI	Public Records Office of Northern Ireland
PWID URI	Uniform Resource Identifier for Persistent Web IDentifiers
ROI	Republic of Ireland
RESAW	Research Infrastructure for the Study of Archived Web Materials
SAA	Society of American Archivists

SPO	State Paper Office
TCD	Trinity College Dublin
TCP/IP	Transmission Control Protocol/Internet Protocol
TEI	Text Encoding Initiative
UK	United Kingdom
UKWA	UK Web Archive
UKWAC	UK Web Archiving Consortium
UNESCO	United Nations Educational, Scientific and Cultural Organization
UN	United Nations
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
US	United States
WAIL	Web Archiving Integration Layer
WARC	Web ARChive file format
WARCnet	Web ARChive studies network researching web domains and events
WARST	Web Archives - Researcher Skills & Tools Survey
XML	EXtensible Markup Language

Glossary

In providing a glossary I direct the reader to a glossary published as part of the WARCnet WARST project, titled 'Towards a Glossary for Web Archive Research: Version 1.0'. An interactive glossary resource is also being developed online in Zotero Groups.

Healy, S., Byrne, H., Schmid, K., Floody, L., Boté-Vericad, J.-J. (2023) *Towards a Glossary for Web Archive Research: Version 1.0*. WARCnet Special Reports. Aarhus, Denmark: WARCnet, https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Healy_et_al_Towards_a_Glossary.pdf_.pdf.

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.

Accessibility

As part of a commitment to accessibility, I have tried to ensure that the URLs provided in the Bibliography and footnotes are (i) captured in a web archive close to the time of access on the live web or (ii) saved in a web archive close to the time of access on the live web. In case of future link rot, I have documented in the Bibliography which web 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.) will also be available to download through the doctoral project files, available in Open Science Framework.[1] Regarding paywall journal articles, I have attempted to provide a DOI, when available, and in the case of open access journal articles, I have further attempted to capture the source URL in a web archive. To further assist with accessibility, I utilise the Arial font for headings, and the Calibri font in the body of the thesis. 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.). I 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 Sharon Healy.

[1] Healy, S., (2022). The Future(s) of Web Archive Research in Ireland [S.C. Healy]. Open Science Framework, https://osf.io/t42va/

ABSTRACT

The central aim of this thesis is to investigate the current state of web archive research in Ireland in line with international developments. Integrating desk research, survey studies, and case studies, and using a combination of research methods, qualitative and quantitative, drawn from disciplines across the humanities and information sciences, this thesis focuses on bridging the gaps between the creation of web archives and the use of archived web materials for current and future research in an Irish context. The thesis describes web archive research to be representative of the web archiving life cycle model (Bragg & Hanna, 2013) which is inclusive of appraisal, selection, capture, storage, quality assurance, preservation and maintenance, replay/playback, access, use, and reuse.

Through a synthesis of relevant literature, the thesis examines the causes for the loss of digital heritage and how this relates to Ireland and explores the challenges for participation in web archive research from creation to end use. A survey study is used to explore the challenges for the creation and use of web archives, and the overlaps, and intersections of such challenges across communities of practice within web archive research. A qualitative survey is used to provide an overview of the availability and accessibility of web archives based in Ireland, and their usefulness as resources for conducting research on Irish topics. It further discusses the influence of copyright and legal deposit legislation, or lack thereof, on their abilities to preserve digital heritage for future generations. An online survey is used to investigate awareness of, and engagement/non-engagement with, web archives as resources for research in Irish academic institutions.

Overall, 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 technologies keep evolving, so too will the challenges. The findings also highlight the need for creators and users/researchers to keep moving forward as collaborators to guide the next generation of web archive research. At the same time, there is also the need for the continual evaluation of legal deposit legislation in line with the fragility of born digital heritage and the technological advances in publishing and communication technologies.

ACKNOWLEDGEMENTS

I have many individuals and organisations to thank for their assistance, support, and encouragement throughout the duration of this PhD project. First, I would like to express my deepest thanks and appreciation to my supervisory team (Joseph, Martin, Juan-José, and Thomas) for their support, encouragement, and guidance over the past few years. My thanks and appreciation also go to the examiners of the thesis, Prof. Susan Aasman and Prof. Thomas O'Connor, who provided insightful feedback which contributed greatly to the overall cohesion of the research. A big shout out to John Sterne of TechArchives, Ireland for his continual kindness in sharing his knowledge, and to Prof. David Malone in the Hamilton Institute (MU) for always finding the time to discuss internet/web history and explain 'techno' stuff to Sharon.

I would also like extend my gratitude to the funders of this research, through the John & Pat Hume Scholarship (Maynooth University), the Government of Ireland Postgraduate Scholarship (Irish Research Council) and the MU-HEA, Covid-19 fund (Maynooth University) for their kind support during a difficult time. The staff of Maynooth University Library must also be thanked, especially the staff dealing with inter-library loans.

I am deeply grateful to the respondents of the survey studies, in giving their time and sharing their experiences, for without them, the full range of this PhD project would not have been possible. I am also very thankful to the WARCnet Steering Group (warcnet.eu) for organising network meetings and activities, which enabled me to learn and network with my peers and experts in the field and participate in collaborative research. Thanks, and appreciation must also go to my collaborators on chapter 4.0 from Maynooth University, the British Library, the International Internet Preservation Consortium, the Bavarian State Library, and the University of Siegen. Special thanks to Helena Byrne from the British Library for her additional collaboration on chapter 5.0. I greatly appreciate her kindness in sharing her knowledge and experience with me over the past few years. To my besties, Bernadette and Lanna, thank you for being there through all the bumps on the road, and for supporting me throughout this journey. Finally, I must thank my mother and children for helping me achieve my goal in attaining a PhD, and yes, I promise, this time I am really finished college!

1.0 INTRODUCTION TO THE RESEARCH

> *The Internet has revolutionized the computer and communications world like nothing before. The invention of the telegraph, telephone, radio, and computer set the stage for this unprecedented integration of capabilities (Leiner et al., 1997, p. 1).*

Since its invention in the early 1990s, the World Wide Web (the web) has become a major resource for researchers (Day, 2003, p. 5; Hendler, 2003, p. 520). Yet it is a transient medium: information is in constant flux with content removal and updates, and the omnipresent '404 Not Found' error. As the early web materialised, concerns about the ephemeral nature of the web also emerged (Brown, 2006, p. 3–4; Pennock, 2013, p. 3). From at least 1994, national libraries and cultural heritage organisations soon realised the need to preserve information and content on the web (Webster, 2017b, pp. 177–178). In the nascent years of the web, there were increasing difficulties for early search engines to index the vast growth of web content through normal cataloguing techniques (Schneider et al., 2009, p. 206; Mirtaheri et al., 2013). Subsequently, specially designed software programs known as web 'crawlers' or 'spiders' started to emerge as a technology to address this from at least 1993. Examples here include World Wide Web Wanderer, Jump Station, and RBSE spider (Mirtaheri et al., 2013). The development of web crawlers also gave rise to the technology for web archiving (Brown, 2006, p. 13; Schneider et al., 2009, p. 206).

There is much consensus that web archiving involves the selection and collection of web content, preserving it for the future and making the collected web content available for access and use (Brown, 2006, p. 5, Dougherty, 2007, p. 19; Niu, 2012; Pennock, 2013, p. 1; Antracoli et al., 2014, p. 157). While a web archive may also come under the umbrella of a digital archive, it is nonetheless, "a specific type of digital archive" (Lomborg, 2019, p. 99). It is worth noting here that there is a difference between web archiving and website backup. Backup software ensures that an organisation's website is copied and retrievable in case of data loss or malfunction. It operates at a more present and recent reference point, as earlier backups tend to be overwritten (Bauer, 2018; Crocetti, 2019). Web archiving on the other hand, is a much more complex process (Bingham & Byrne, 2021, p. 3; Antracoli, 2014, p. 157; Brügger, 2018, p. 81) and is representative of the processes and activities described in the Archive-It web archiving lifecycle model which is inclusive of appraisal, selection, capture,

storage, quality assurance, preservation and maintenance, replay/playback, access, use and reuse (Bragg & Hanna, 2013) (Figure 1.1).

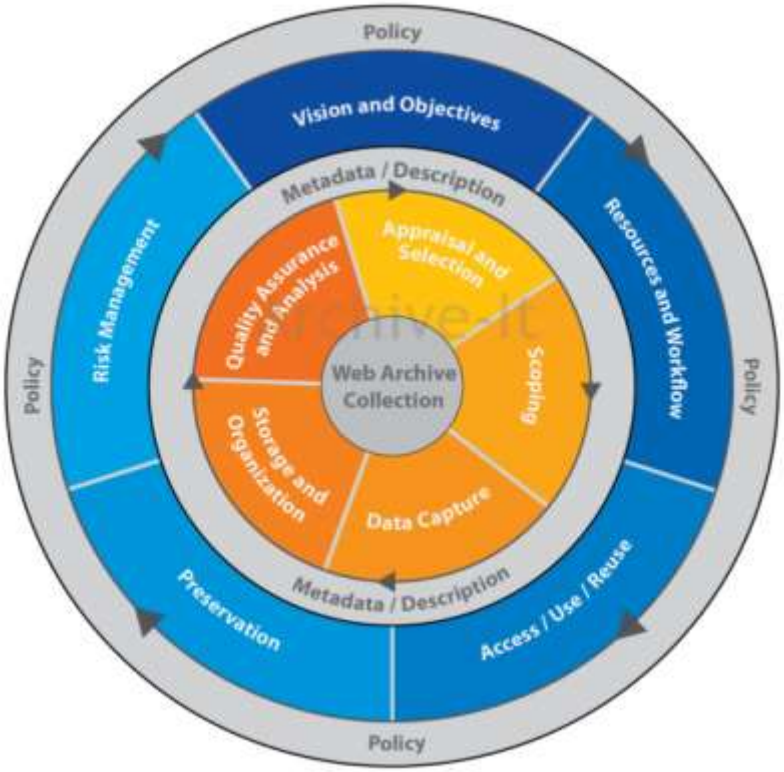

Figure 1.1: Web Archiving Life Cycle Model (Bragg & Hanna, 2013) which is inclusive of appraisal, selection, capture, storage, quality assurance, preservation and maintenance, replay/playback, access, use and reuse

In addition, the terms 'internet' and 'web' are often used interchangeably, and while they are connected, they are separate entities. In brief, the internet is a networking infrastructure which connects computers, devices, and mobile phones on a global scale, whereas the web is an interlinked information system, operable through the medium of the internet (Beal, 2010; Milligan, 2019, p. 32).

In describing the difference between the internet and the web, Gillies and Cailliau (2000) offer a useful explanation suggesting that the internet is

> like a network of electronic roads criss-crossing the planet - the much-hyped information superhighway. The Web is just one of the many services using that network, just as many different kinds of vehicles use the roads (p. 1).

The internet, as we know it today, for the most part operates via Transmission Control Protocol/Internet Protocol (TCP/IP). First, an IP address is an allocated number for a machine connected to the internet, whether it is a laptop, an X-box, or a smart TV. Milligan (2019)

suggests thinking of an IP address as a library call number, “letting us quickly locate things in an otherwise overwhelming sea of items” (p. 36). In explaining TCP/IP further, Computer Hope provides the following explanation.

> TCP/IP is a set of rules (protocols) governing communications among all computers on the Internet. More specifically, TCP/IP dictates how information should be packaged (turned into bundles of information called packets), sent, and received, as well as how to get to its destination (Computer Hope, n.d.).

In the Republic of Ireland, the first public connection to the internet (via TCP/IP) went live in Trinity College Dublin in June 1991. University College Dublin also connected several weeks later (Sterne, 2015+). In December 1994, President Mary Robinson became the first Irish Head of State to make use of email “sending Christmas greetings to thousands of emigrants via the Internet” (Cunningham, 1995, p. 17). In February 1995, the Department of Foreign Affairs became the first Irish government department to disseminate an official document via the web, albeit from “a server in the computer applications department at Dublin City University” (Sterne, *IT’s Monday*, 1995: Issue 142).

Today, the web is accessed via browsers such as Google Chrome, Microsoft Edge, Firefox, or Safari, and allows for access to web pages which contain a collection of resources and files such as: text, images, graphics, books, newspapers, audio, video, movies, databases, widgets, styling, scripts, software and more (Milligan 2019, p. 4; Day, 2006, pp. 17–18). According to Brügger (2018), the web can be examined through an analytical grid of five strata: an individual web element, an individual web page, an individual website, a web sphere, and the web in its entirety (p. 31). Moreover, this can be applied equally to both layers of the web being “the visible/audible web in the browser, and hidden text of HTML code and associated files” (Brügger, 2018, p. 31). These five strata also offer an equally applicable model for studying the archived web.

For Brügger (2018) there are several types of “active processes” which we might consider to be a form of web archiving as follows:

1. making an image like a screenshot PNG or PDF,
2. making a screen movie or screencast, for example recording a user “moving through a website” or playing a video game,
3. downloading individual files like HTML files, or text extracted from HTML files, or embedded files on a specific web page such as images, audio, or video,
4. web crawling which tends to be used by institutions invested in the large-scale preservation of websites,

5. collecting material through APIs like social media,
6. collecting websites that were taken offline but are preserved intact, for example, a backup of a website stored on a media carrier such as a floppy disk or other types of media carriers,
7. collecting web materials as they appear in other types of media such as books, documentaries, television, or film. Although here Brügger notes that this may not really be classified as there is "no active process taking place" with the content, but worth including as it may be the only source available, especially from the early years of the web (Brügger, 2018, pp. 80–81).

Social media archiving also comes under the umbrella of web archiving. For Thomson (2016), social media

> plays an increasingly important role as we embrace networked platforms and applications in our everyday lives. The interactions of users on these web-based platforms leave valuable traces of human communication and behaviour revealed by ever more sophisticated computational analytics. This trace – the data generated by social media users – is a valuable resource for researchers and an important cultural record of life in the 21st century (p. 1).

Therefore, Thomson (2016) suggests that social media archiving acts as a support for research, public records, and cultural heritage. Thomson (2016) further proffers that as "social media continues to grow as a source of official government and corporate communications, the importance of effective preservation will increase" (p. 1).

On the other hand, it should also be noted that social media is a challenging online format to archive. The workflows and tools used to capture social media often necessitate a different approach to archiving static or semi-static pages through web crawling and may necessitate collecting data through Application Programming Interfaces (APIs). Byrne (2017) describes how archiving social media can be "technically challenging" due to the fact that some platforms are specifically designed to prevent access to crawlers. Nevertheless, there are several strategies that can be used to collect social media content, in some (but not all) cases. Collecting and preserving social media also present different legal, ethical, and curatorial considerations (Thomson, 2016; Breed, 2019; Bingham et al., 2020; Bingham & Byrne, 2021; Michel et al., 2021; Vlassenroot et al., 2021). For example, Thomson (2016) discusses how social media "requires particular solutions for access, curation, and sharing that accommodate the particular curatorial, legal, and technical frameworks of large aggregates of machine-readable data" (p. 4). Nonetheless, social media content is increasingly being identified as records which have long-term preservational value. For

instance, the National Archive of Australia insists that “Social media and instant messaging posts, media, comments, messages and analytics that are created or received as part of Australian Government business are Commonwealth records” (National Archives of Australia, n.d.). Also, as part of the UK Government Archive service, The National Archives, UK archives “social media output of government as it represents an important part of how government communicates online” (Storrar, 2018). It might also be noted that future historians of the COVID epidemic (2019-2022) will be reliant on access to social media records for any analysis of its impact on communities.

While there are several ways in which the web or social media might be archived, and while it is important to acknowledge such efforts, the main focus of this thesis is on institutional web archiving through web crawling, and the use of institutional web archives for research or other purposes.

1.1 Web Archive Research

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). Although, coming up with a universal definition for web archive research is not the goal of this thesis, there is a need for 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 it 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) description of web archive research as "the study of all activities involving web archives", fits well for the purpose of this doctoral research. However, the thesis also considers 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, selection, capture, storage, quality assurance, preservation and maintenance, to replay/playback, access, use and reuse (Bragg & Hanna, 2013). In addition, as noted previously, the focus of this thesis, for the most part, is on institutional web archiving through crawling, and the use of institutional web archives for research or other purposes.

Web archiving, through 'crawling', uses an archival quality web crawler to retrieve and save content from the web through a process called 'capturing' or 'harvesting'. For the most part archival quality crawlers have been developed by the web archiving community themselves. For example, the Heritrix crawler was developed by the Internet Archive from early 2003. With cooperation from the Nordic national libraries in the latter part of 2003, Heritrix had its first public release as open source in January 2004 (Mohr et al., 2004, p. 4). Private web archiving companies may also develop their own in-house crawlers. For instance, the company MirrorWeb uses Heritrix, but also developed Electrolyte as a crawler to explore and capture "the most rigorous and dynamic digital domains" (MirrorWeb, n.d., SEC 17a-4). Beginning from an initial set of identified Uniform Resource Locators (URLs), known as 'seed' and/or 'target' URLs , the crawler contacts the web server with a request to retrieve and save the pages that are identified by the URL. The crawler then finds all the hyperlinks on a page that link to other pages and files such as images, PDFs, styling sheets, scripts etc., and lists these URLs to the crawl queue. The process is repeated until the queue is emptied, or it reaches a specified URL threshold limit (International Internet Preservation Consortium, n.d., The WARC Format 1.0; Brügger, 2018, p. 81; Mirtaheri et al., 2013; Cho & Garcia-Molina, 2000, p. 200).

Capturing content on the 'surface' web presents problems due to the wide range of data types such as text, image, sound, visual, multimedia and even software in varied formats "all of which may need to be considered separately from a preservation perspective" (Day, 2006, pp. 17–18). The 'surface' web can be accessed by a URL, whereas the 'deep' Web is estimated to be far larger, and access often requires encryption keys or an authentication log-in and password, or requires payment for access (Lyman, 2002, p. 41; Day, 2003, pp. 16–17). Bergman (2001) notes how traditional search engines by-pass or cannot retrieve some deep web content as some web pages "do not exist until they are created dynamically as the result of a specific search." Indeed, to view information on the 'deep' web often involves user interaction and the input of a request (Bergman, 2001).

The captured data also needs to be filed and stored. Song and Jaja (2008) suggest that there are a few methods used, but the most popular method for institutional archiving is via "containers in a well-defined structure" such as ARC/WARC formats (p. 2). The Internet Archive ARC file format (ARC_IA) is a method for merging multiple digital resources together (e.g., HTML, CSS, JPG, PNG etc.) into a self-contained aggregate archival file together with related information (Library of Congress, n.d., ARC_IA; Burner & Kahle, 1996). Most institutions now prefer the WARC (Web ARChive) file format, which is an extension of the ARC format with revisions to accommodate "related secondary content, such as assigned metadata, abbreviated duplicate detection events, later-date transformations, and segmentation of large resources", although, ARC is still an acceptable legacy format (International Internet Preservation Consortium, n.d., The WARC Format 1.0).

The captured data is then processed to be part of a web archive collection, where access is provided through replay or playback software which offers some form of search interface such as the Wayback Machine, or its open-source counterpart, OpenWayback (Costa, 2021, p. 73). The use and reuse of the captured data completes the sequence of the full web archiving lifecycle. The use of the archived web can be found across various fields of research, covering topics such as internet and web histories, sport history, social, health and political sciences, information sciences, law, media and journalism, religious discourse online, youth and social justice, diasporas, migrant communities, and more (Raffal, 2018; Aasman, 2019; Kahn, 2019; Milligan, 2019; Byrne, 2019; Adelmann & Franken, 2020; Gorsky, 2015; Ben-David, 2019; Foot & Schneider, 2006; Cocciolo, 2015; Holzmann et al., 2016; Eltgroth, 2009; Taylor, 2017b; Weber & Napoli, 2018; Bødker & Brügger, 2018; Hofheinz, 2010; Webster, 2019; Webster, 2017a; MacKinnon, 2020; MacKinnon, 2021; Huc-Hepher & Wells, 2021).

Studies which offer examples for the reuse of web archive data are difficult to find, with some exceptions being Ruest (2016), Sherratt and Jackson (2021), Brügger (2021a, 2021b), and Eldakar and Holownia (2022). The reuse of data from web archives is problematic for several reasons, some of which include 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).

Finally, the lifecycle of the captured data also includes active maintenance to ensure its long-term digital preservation. The American Library Association describes digital preservation as a combination of

> policies, strategies and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time (American Library Association, 2008).

Hence, for many commentators, web archiving is a "complex" process (Bingham & Byrne, 2021, p. 3; Antracoli, 2014, p. 157; Brügger, 2018, p. 81) which requires a great deal of decision making.

Web archiving decisions must be made on the selection of content to be captured; the technology to use for capturing, storage, preservation, and replay/playback; as well as how to make the collected data accessible for use, and indeed, how flexible this access might be. Furthermore, such decisions may be influenced by social, cultural, and political circumstances; legislations on copyright and legal deposit or lack thereof; and the availability of resources in terms of finance, labour, technology, and organisational infrastructures (Ogden, 2021; Dougherty, 2007; Ben-David, 2019; Ben-David, 2021; Hockx-Yu, 2014; Winters, 2020a; Winters, 2019; Vlassenroot et al., 2019; Brügger, 2021c; Maemura, 2022). For Vlassenroot et. al. (2019)

> 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).

Thus, as more of the cultural, historical, legal, evidential, informational, and social record happens on the web, heritage institutions are tasked with keeping up with ongoing technological changes to capture and preserve this transient medium.

1.2 Research Problem

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; Ben-David, 2021, p. 181). 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; Gomes et al., 2021a). Indeed, despite the increasing availability of archived web content, scholars have highlighted how academics have been slow to embrace web archives as resources for research (Webster, 2020; Rogers, 2019; Leetaru, 2019; Meyer et al., 2017; Webster, 2017b; Winters, 2017; Leetaru, 2017; Brügger, 2016; Meyer et al., 2011; Dougherty et al., 2010). For example, Meyer et al. (2011) believe that "the use cases for web archives are not well articulated and have not engaged

the research community in any significant way" (p. 4). Dougherty et al. (2010) conclude that there is

> a gap between the potential community of researchers who have good reason to engage with creating, using, analysing and sharing web archives, and the actual (generally still small) community of researchers currently doing so (p. 5).

Truman (2016) identifies the need for more communication and collaboration between those who curate, create, and steward web archives, and those who use (or might use) a web archive for purposeful research (p. 3).

In relation to Ireland, publication of Irish based research integrating the use of archived web content is difficult to find with a few exceptions being Malone (n.d.), Harjani (2018), Byrne (2019), Webster (2019), Greene & Ryan (2019), and Greene (2020). Also, at the time of this research, and to the best of my knowledge, there have been no web archive user studies conducted across Irish academia that examine scholarly engagement, or awareness of the existence of web archives as resources for research. Moreover, as pointed out by the web archivist at the National Library of Ireland, "'It's difficult to get good analytics on web archive users, due to the fact [that] the selective web archive can be accessed remotely'" (Ryan cited in Vlassenroot, 2019, p. 100). In essence, very little is known about those who engage with, or might potentially engage with, web archives as resources for Irish based research. Clearly, there is a void between the creators of web archives and the users or potential users of these archives. This thesis is concerned with bridging the gaps between the creation of web archives and the use of archived web materials for current and future research within an Irish context.

1.3 Thesis Aims

The central aim of this thesis is to investigate the current state of web archive research in Ireland in line with international developments. Integrating desk research, survey studies, and case studies, and using a combination of research methods, qualitative and quantitative, drawn from disciplines across the humanities and information sciences, the thesis focuses on bridging the gap between the creation of web archives and the use of archived web materials for current and future research in an Irish context.

First, the thesis positions heritage within a wider framework of societies, communications, and culture, as it is within the intersections of these concepts that heritage is produced. This will facilitate a deeper understanding of why societies and communities feel the need to preserve and pass on their heritage, in the first place, and foster a better understanding of

digital heritage. The thesis then seeks to identify the reasons for the loss of digital heritage and how this relates to Ireland. It further aims to examine skills, tools, and knowledge ecologies within web archive research on an international level and explores the challenges for the creation and use of web archives, as well as the overlaps and intersections of such challenges across communities of practice within web archive research. The thesis also seeks to provide an overview of the landscape of web archives based across Ireland, and their availability and accessibility as resources for Irish based research. Additionally, the thesis investigates the awareness of, and engagement/non engagement with, web archives as resources for research in Irish academic institutions. Finally, the thesis aims to explore ways in which to improve the conditions for conducting web archive research, and how this relates to Ireland.

For clarity, this thesis refers to Irish digital heritage in the context of the digital heritage of the island of Ireland. When required, it will refer to the digital heritage of Northern Ireland or the Republic of Ireland to distinguish between the two jurisdictions.

1.3.1 Research Questions

In pursuit of the aims above, this thesis is guided by the research questions outlined in Table 1.1.

Table 1.1 Research questions in line with thesis chapters

RQ1:	What are the main causes for the loss of digital heritage? How does this relate to Ireland?	Chapter 2.0 Chapter 5.0
RQ2:	What are the main challenges for participation in web archive research? How does this relate to Ireland?	Chapter 2.0 Chapter 3.0 Chapter 4.0 Chapter 5.0 Chapter 6.0
RQ3:	How available and accessible are web archives based on the island of Ireland for conducting Irish based research?	Chapter 5.0
RQ4:	What is the current level of awareness of, and engagement/non-engagement with web archives in Irish academic institutions?	Chapter 6.0
RQ5:	How can we improve the conditions for conducting web archive research, and how does this relate to Ireland?	Chapter 2.0 Chapter 3.0 Chapter 4.0 Chapter 5.0 Chapter 6.0

1.4 Structure and Methodologies

The thesis integrates desk research, survey studies, and case studies, and uses a combination of research methods, qualitative and quantitative, drawn from several discipline areas. Primary data was collected through an online web archive user survey, and through another online survey with a focus on individuals who participate in web archive research. In a broad sense, the thesis may be positioned at the intersection of the humanities and information science. Within this, the thesis engages with scholarship and perspectives from archival science, library and information science, heritage studies, computer science, social sciences, media studies, cultural studies, humanities studies, and the evolving field of web archive research. Each chapter of the thesis offers a distinctive component of the research and provides a standalone research design and methodology. Therefore, there is no need to duplicate those efforts here. However, the following section offers a brief overview of each chapter in line with a summary of the chapter methodologies.

Using desk research and a literature review from across multiple disciplines (as mentioned above), Chapter 2.0 is concerned with examining the preservation of tangible heritage through the activities of archives and libraries, with a particular focus on the preservation of national digital heritage. First, it positions heritage within a wider framework of societies, communications, and culture, as a first step in formalising an understanding of how heritage is produced, and thus, enabling a deeper understanding of the causes for the loss of digital heritage, including Irish digital heritage (RQ1). The chapter explores some of the underlying reasons for web archiving, and how they stem from wider concerns on the loss of digital heritage in general. The chapter also addresses (RQ2), through a discussion on the advances in electronic publishing and communications technologies, and how nation states need to review and amend their legislations regarding copyright and legal deposit, to be inclusive of web archiving. The literature also provides some insights for the improvement of conditions for conducting web archive research (RQ5).

Chapter 3.0 similarly uses desk research and a literature review from across multiple disciplines to provide an overview of web archive research. It explores some of the practices and principles for web archiving, as well as some of the challenges (RQ2), and examines scholarly engagement with web archives, and the challenges experienced by this user community (RQ2). A brief overview of the literature relevant for studying the archived web is presented. This is followed by a discourse on the value of web archives for research or other purposes. The final section provides an overview of the literature related to the thesis in line with studies on web archiving practices and tools, as well as web archive users and

scholarly engagement. This literature also offers some research and perspectives that would be useful for improving the conditions for conducting web archive research (RQ5).

Through a collaborative interdisciplinary project (WARST) Chapter 4.0 examines the challenges for participation in web archive research on an international level (RQ2) and offers some approaches which would be useful for improving the conditions for conducting web archive research (RQ5). The chapter engages with research methods within information sciences for the collection and analysis of quantitative and qualitative data through a survey study in order to explore the skills, tools, and knowledge ecologies in web archive research and the challenges for participation. In doing so, it 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. The chapter seeks to identify and document skills, tools and knowledge ecologies within the web archiving lifecycle and explores the challenges for participation in web archive research and the overlaps and intersections of such challenges across communities of practice. The chapter also provides some useful perspectives for improving the conditions for conducting web archive research (RQ5).

Through another collaborative effort, Chapter 5.0 uses a qualitative exploratory approach through desk research, a review of the literature, and informal dialogues with heritage colleagues to examine the availability and accessibility of web archives based on the island of Ireland, as well as their usefulness as resources for Irish based research (RQ3). The chapter also examines the causes for the loss of Irish digital heritage (RQ1) and discusses the challenges for web archive research in the context of Irish digital heritage (RQ2). Presently, there are three main web archiving initiatives which capture and preserve websites as part of their efforts for the preservation of digital heritage for the island of Ireland. These are: the PRONI Web Archive, the UK Web Archive, and the NLI Web Archive. The chapter offers an overview of these web archiving initiatives and their historical backgrounds, inclusive of the influences of copyright and legal deposit legislation on their collecting activities. The chapter further examines their efforts for the collection and preservation of digital heritage from the web spaces of Northern Ireland and the Republic of Ireland and assesses their availability and accessibility as resources for research on Irish based topics. Some insights on approaches which would be useful for improving the conditions of web archive research (RQ5) in an Irish context are also unearthed.

Chapter 6.0 investigates the levels of awareness of, and engagement with, web archives in Irish academic institutions (RQ4). It also examines the perceived challenges by Irish based

researchers for the future use of archived web content in their field of study (RQ2). The chapter engages with a review of related literature from across disciplines, but with a focus on web archive user studies, scholarly engagement with web archives, and literature which uses web archives for research on Irish based topics. It further uses an online survey to investigate awareness of web archives, and engagement/non-engagement with web archives by lecturers, researchers, and students in Irish academic institutions. The chapter also addresses (RQ5) by offering some thoughts on approaches which would be useful for improving the conditions of web archive research.

Through a synthesis of the findings and discussions, Chapter 7.0 revisits the research problem and the research questions and answers. It concludes with some final thoughts and suggestions for future work.

1.5 Research Contribution

This thesis contributes to the international literature by offering a detailed analysis of the challenges faced by both the creators and users of web archives and how these challenges overlap across communities of practice within web archive research. In doing so, it reinforces the importance of collaborations between web archive creators and users as a necessary component for the future development of web archive research and offers suggestions and approaches for improving the conditions for web archive research across communities of practice. The thesis is timely, as it will contribute to the current debates in the Republic of Ireland regarding the necessity for the implementation of legal deposit legislation which realistically reflects the fragility of born digital heritage and the technological advances in publishing and communication technologies.

To the best of my knowledge, there appears to be no known web archive user studies conducted across Ireland which examine scholarly engagement, or awareness of the existence of web archives as resources for research. Literature on Irish based research integrating the use of archived web content is difficult to find except for Malone (n.d.), Harjani (2018), Byrne (2019), Greene & Ryan (2019), Healy (2019), Webster (2019), and Greene (2020). In essence, very little is known about those who engage with, or who might potentially engage with, web archives as resources for Irish based research. There is no doubt therefore that the outcome of this study would help to fill up this void. The thesis generates awareness for the loss of Irish digital heritage, while making a case for the urgent need to implement digital preservation strategies in the Republic of Ireland for the preservation of electronic records, multimedia and born digital materials.

1.6 Collaborations

Some of the research for this thesis was conducted through collaborative interdisciplinary work and is accounted for next.

Chapter 4.0 in its entirety encompasses work from a research project titled Web Archives – Researcher Skills & Tools Survey (WARST). WARST is a collaborative interdisciplinary 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, with backgrounds in humanities, digital humanities, cultural studies, media studies, cultural heritage, library and information science, archival science, computer science, and IT development.

Chapter 5.0 also encompasses work from a collaborative study by Sharon Healy (Maynooth University) and Helena Byrne (British Library). The study assesses the availability and accessibility of web archives which would prove useful as resources for Irish based research. Therefore, it was essential to apply both a researcher-centric and curatorial-centric approach to this study, and a collaboration was required.

1.7 Dissemination

1.7.1 Publications

Healy, S. & Byrne, H. (2023). *Scholarly Use of Web Archives Across Ireland: The Past, Present & Future(s)*. WARCnet Special Report. Aarhus, Denmark: WARCnet, https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Healy_Byrne_Scholarly_Use_01.pdf. [URL Memento: Wayback Machine]

Healy, S., Byrne, H., Schmid, K., Floody, L., Boté-Vericad, J.-J. (2023) *Towards a Glossary for Web Archive Research: Version 1.0*. WARCnet Special Report. Aarhus, Denmark: WARCnet, https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Healy_et_al_Towards_a_Glossary.pdf. [URL Memento: Wayback Machine]

Healy, S., Byrne, H., Schmid, K., Bingham, N., Holownia, O., Kurzmeier, M., & Jansma, R. (2022). *Skills, Tools, and Knowledge Ecologies in Web Archive Research*. WARCnet Special Report. Aarhus, Denmark: WARCnet, https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Healy_et_al_Skills_Tools_and_Knowledge_Ecologies.pdf. [URL Memento: Wayback Machine]

Healy, S. (2019). Web archives as resources to find archived treasures. MU Library Treasures, 30 November 2019, https://mulibrarytreasures.wordpress.com/2019/11/30/web-archives-as-resources-to-find-archived-treasures/. [URL Memento: Wayback Machine]

Forthcoming

Byrne, H., Boté-Vericad, J.-J. & Healy, S. (2024 forthcoming). Skills & Training Requirements for the Web Archiving Community. In Brügger, N. et al. (Eds.) *The Routledge Companion to Transnational Web Archive Studies*. London, New York: Routledge.

1.7.2 Conference Presentations & Posters

Healy, S., Byrne, H., Schmid, K., Bingham, N., Holownia, O., Kurzmeier, M., & Jansma, R. (2022). An Overview of Skills, Tools and Knowledge Ecologies in Web Archive. *WARCnet Closing Conference, Aarhus University, Denmark, 17-18 October 2022*, You Tube: https://www.youtube.com/watch?v=yf9GdGSzob4&ab_channel=UKWebArchive.

Schmid, K., Healy, S., Byrne, H. (2022). Exploring Software, Tools and Methods used in Web Archive Research. *iPres 2022: International Conference on Digital Preservation, Glasgow, Scotland, 12-16 September 2022*, https://bl.iro.bl.uk/concern/conference_items/4943dae4-fbd5-40fa-85a6-99e63638bee0?locale=en. [British Library Repository]

Healy, S. (2022). Web Archives as Liminal Spaces. *Society and the Arts in the Pandemic, Symposium, Dundalk Institute of Technology, Ireland, 29 April 2022*.

Healy, S., Holownia, O., Kurzmeier, M., Webber, J. (2021) Introducing the Web Archives – Researcher Skills & Tools Survey (WARST). *Engaging with Web Archives for Digital Humanities, Maynooth University Arts and Humanities Institute, Co. Kildare, Ireland, 01 September 2021,* https://ewaconference.com/ewa4dh-2021/ewa4dh-programme/. [URL Memento: Wayback Machine]

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

Healy, S. (2019). Coming Out in Éire: exploring methodologies for finding and recording internet and web histories of the LGBT movement in Ireland. *RESAW '19 – 'The Web that Was', University of Amsterdam, the Netherlands, 19-21 June 2019*. Research Infrastructure for the Study of Archived Web Materials, https://easychair.org/smart-program/RESAW19/2019-06-21.html#talk:89177. [URL Memento: Wayback Machine]

Healy, S. (2017). The web archiving of Irish election campaigns: A case study into the usefulness of the Irish web archive for researchers and historians. *RESAW/IIPC Conference, School of Advanced Study, University of London, 14-16 June 2017,* https://easychair.org/smart-program/RESAW19/2019-06-21.html#talk:89177. [URL Memento: Wayback Machine]

Healy, S. (2016). Here today, gone tomorrow: A case study on the necessity for a more rigorous approach to the preservation of online Irish cultural and political heritage. *Institutions and Ireland: Public Cultures, Trinity College Dublin, 27 October 2016,* DOI: 10.17613/xy5v-6b63.

1.7.3 Zotero Resources

Healy, S. (2022). Zotero Groups - The Future(s) of Web Archive Research Across Ireland. Zotero, https://www.zotero.org/groups/4712321/the_futures_of_web_archive_research_across_ireland_s.c._healy.

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.

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.

1.7.4 Open Science Framework (OSF) Resources

Healy, S., Byrne, H., Schmid, K., Bingham, N., Holownia, O., Kurzmeier, M., Jansma, R., Jane Winters, & Boté-Vericad, J.-J. (2022). Skills, Tools, and Knowledge Ecologies in Web Archive Research (WARST Project). Open Science Framework, https://osf.io/vf7gt/.

Healy, S., Byrne, H. (2023). Scholarly Use of Web Archives Across Ireland: The Past, Present, and Future(s). Open Science Framework, https://osf.io/crfnw/.

Healy, S., (2022). The Future(s) of Web Archive Research Across Ireland. Open Science Framework, https://osf.io/t42va/.

2.0 RECOGNISING THE PROBLEMS

Throughout time individuals and societies have communicated, captured and passed on many of their stories by selectively storing, structuring, and re-presenting them - graphically, textually, on some kind of media and using whatever technology is available to them - the chalk on the cave wall, the carving on the monolith, the paint on the clay pot or the mummy case, the handwriting on the scroll, the sound recording on the CD, the bits on the computer disk, the image on the film. Other stories are remembered by being told, sung, danced or performed, captured in rituals and ceremonies, recalled and retold or performed again (McKemmish, 2005, p. 2).

2.1 Introduction

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). By the early 2000s, the web was claimed to be "the information source of first resort for millions of readers" (Lyman, 2002, p. 38). However, 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.

Over the past decades, 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; Cho & Garcia-Molina, 2000; Dellavalle et al., 2003; Ntoulas et al., 2004; Sellitto, 2005; Goh & Ng, 2007; Wren, 2008; Klein et al., 2014; Zhou et al., 2015; Bansal & Parmar, 2020; Craigle et al. 2022). Concerns about the ephemeral nature of the web also stemmed from existing apprehensions regarding the storage and preservation of computerised records, electronic information, multimedia and born digital materials in general (Fishbein, 1972; Dollar, 1978; Committee on the Records of Government, 1985; Graham, 1994; Waters & Garrett, 1996; Gardner, 1997; Kuny, 1997). Web archiving has emerged as a means for collecting and preserving born digital heritage from the web.

Using desk research and a literature review, this chapter is concerned with examining the preservation of tangible heritage through the activities of archives and libraries, with a particular focus on the preservation of digital heritage. It engages with research and perspectives from archival science, library and information science, heritage studies, computer science, social sciences, media studies, cultural studies, humanities studies, and the evolving field of web archive research, to examine some of the main causes for the loss of digital heritage, and how this relates to Ireland (RQ1). First, the chapter positions heritage within a wider framework of societies, communications, and culture, for it is within the intersections of these concepts that heritage is produced. This will facilitate a deeper understanding of why societies and communities feel the need to preserve and pass on their heritage, in the first place, "using whatever technology is available to them" (McKemmish, 2005, p. 2). The chapter explores some of the underlying reasons for web archiving, and how this stems from wider concerns on the loss of digital heritage in general (RQ1). It discusses the advances in electronic publishing and communications technologies, and how this resulted in the need for nation states to review and amend their legislations regarding copyright and legal deposit, to be inclusive of web archiving (RQ2). The literature also provides some insights for the improvement of conditions for conducting web archive research (RQ5).

The Irish Heritage Council (n.d.) describes heritage as "what we have inherited from the past, to value and enjoy in the present, and to preserve and pass on to future generations." This includes tangible heritage (sites, monuments, artefacts, archives, libraries, museums, etc.), natural heritage (landscapes, waterways, native plants, wildlife, insects etc) and intangible heritage (customs, sport, music, dance, traditions, myths) (The Heritage Council, Ireland, n.d.). It is the combination of these three strands of heritage which "provide us with a common language and insight that enables us to communicate on a deep level with each other and to express ourselves in a unique way to the outside world" (The Heritage Council, Ireland, n.d.).

Therefore, heritage institutions such as galleries, libraries, archives, and museums (GLAM) play important roles in the preservation of societal heritage. The UNESCO (2003) Charter on the Preservation of the Digital Heritage, characterises digital heritage as "unique resources of human knowledge and expression" that embrace

> cultural, educational, scientific and administrative resources, as well as technical, legal, medical and other kinds of information created digitally, or converted into digital form from existing analogue resources. Where resources are 'born digital', there is no other format but the digital object (UNESCO, 2003).

The charter further states that “the disappearance of heritage in whatever form constitutes an impoverishment of the heritage of all nations”, and so, digital heritage should not be an exception (UNESCO, 2003).

2.1.1 Archives and Libraries

For some commentators the history of the archive can be traced to the ancient world of Mesopotamia in the Near East, and the findings of large collections of Sumerian and Akkadian clay tablets, described as “administrative records” from “various centres of the realm of the Third Dynasty of Ur” around 2100 B.C. (Posner, 1972; Veenhof, 1983). Both Posner (1972) and Veenhof (1983) discuss whether the finding of clay tablet collections might best be described as belonging to an archive or library, and while both terms have been applied interchangeably, there are cases where the collections are a mixture of both. Veenhof (1983) describes how a collection of tablets used as literary and scientific texts (e.g., “school texts”) and tablets used for administrative and economic records were sometimes stored together (pp. 5–6). The histories of libraries are often traced to the clay tablet collections belonging to the royal library of Ashurbanipal 700 B.C. who reigned over the Assyrian Empire in Mesopotamia, mainly because it contained a large body of clay tablets with texts, nonetheless, it also contained some tablets which are best described as palace records and legal documents (Veenhof, 1983, p. 6). Further findings in the administrative quarter of the Royal Palace G of Ebla, a city state in ancient Mesopotamia would reveal a “Central Archive” of clay tablets dating from 2400 B.C. to 2250 B.C. mainly related to administrative records of economic/commercial activities, while some clay tablets were literary texts (Archi, 2015; Bradsher, 2020).

Prior to the invention of paper, clay tablets were not the only media format used for recording literary texts, administrative records, and legal documents. Papyrus, leather, wax, wood, and later parchment were also used for such purposes. However, these media types were more perishable due to fire, flooding, or climate factors and are rarely found outside of places with hot climates, such as the papyrus found in the Egyptian desert (Veenhof, 1983). Because of this, Veenhof (1983) states that historians of Mesopotamia are in a more favourable position in studying ancient civilisations “compared to scholars studying countries and civilizations which used papyrus, leather, parchment or paper for daily recording” (p. 2). Veenhof (1983) argues that

> Even when rich epigraphical remains are available, like from ancient Egypt, one is faced with the effects of a "natural" selection, since as a rule only ceremonial inscriptions on stone etc. and texts deposited in places where destruction and

> climatic influences had little effect (such as tombs in the desert) have survived, while the bulk of what was written for administrative purposes has perished. The contrast is obvious in places where both clay tablets and papyri were written and kept, such as El Amarna, where only the official correspondance on clay was discovered (Veenhof, 1983, p. 2).

Both the Hellenistic (365 to 30 B.C.) and Roman republic governments (509 to 27 B.C.) established archives as a place for the preservation of public records that were available for consultation (Posner, 1972; Erskine, 2009). The Hellenistic period also contributed to the establishment of libraries for public learning, such as the Library of Alexandria in Egypt which came into being during the reign of Ptolemy II (283 to 246 B.C.) but was influenced in its establishment by (his father) Ptolemy I, and his private collection of texts (Tracy, 2000, p. 343–344). The collection included texts from the Peripatetic school of philosophy, which was founded by Aristotle in Ancient Athens in 335 B.C. (Tracy, 2000, p. 344).

Archives were also established through the evolution of city states and then the evolution of state formation across Europe. The demise of the feudal systems, the development of towns, and the evolution of the concept of "republican" governments, especially in the Italian city states from the seventh century onwards were important factors in the development of the modern state (Nelson, 2006, p. 55). The purpose of city state archives was for public and political administration. Thus, the maintenance and security of the archive, as an administrative memory, became a responsibility of city states, and later the modern state. For Yale (2015) archives played "key roles in the formation and governance of nation states and empires" and may be seen as "instruments through which political power was (and is) exercised" (p. 336). As a result of the Thirty-Year War, a state system in Europe was recognised in the treaty of Westphalia (1648) and the concept of the modern state "finally emerges in clear form" as sovereignty coincided with territory (Nelson, 2006, p. 60).

Other institutions, such as the Papacy or the monastic centres, also began to create their own archives in imitation of city state practices. Pope Innocent III was one of the first to formalise the church's archival policy during his reign at the end of the twelfth century. However, many records were lost as the archival records tended to move with the Pope as he travelled. In 1565, Pope Pius IV set about creating a centralised church archive in the Vatican Palace to make access easier for administrators and is credited with the making of the modern-day Vatican Archives (Coombs, 1989). Moreover, from the sixth century onwards, monastic centres were also involved with the copying and preservation of literary texts, and the adoption of parchment as a media format for copying manuscripts (Muldoon,

1997). This stage came about due to concerns for the preservation of "the classical and Christian literary heritage of the Roman world" which was written on papyrus (p. 49). Papyrus was only suited for the warm dry region of Egypt, but "decayed rapidly" in the cold and wet climates "north of the Mediterranean" (p. 49). This copying process "not only served to retain this ancient knowledge, but to spread it as well, first as monasteries sought to expand their libraries by obtaining copies of manuscripts found elsewhere" (Muldoon, 1997, p. 49). For Muldoon (1997), the copying of manuscripts on parchment would then be disrupted by the invention of the printing press with moveable type from the mid-fifteenth century onward.

While Muldoon (1997) acknowledges the popular belief that Johann Gutenberg lay at the helm of the invention, he notes how "Gutenberg was only one of several individuals who were experimenting with printing" and how there was "a widespread interest in producing books by technological means rather than by hand" (p. 50). Muldoon (1997) further discusses how a printing press with movable type would have been of little use, had it not been for the invention of paper and the alphabet beforehand (p. 50). Similarly, McLuhan (1962) states that "without the alphabet, there would have been no Gutenberg" (p. 40). The successful development of printing also relied "upon other improvements such as those in metallurgy--the discovery of the proper mix of lead, tin, and antinomy to be used in the casting of type" (Muldoon, 1997, p. 51). For Muldoon (1997), it was "the bringing together of several developments, the synthesizing so to speak of several technologies in order to achieve the final product, and then subsequent refinement of the result" (Muldoon, 1997, p. 51). Muldoon (1997) also points to the popular myth of how the printing press increased literacy across Europe – he suggests that this negates the fact that there were already some good levels of literacy in urban Europe, the Rhineland, the Netherlands, and Italy, and "the market demand for the written word pushed the development of printing rather than the reverse" (p. 51).

The establishment of modern archival practices tends to be correlated with the decrees of the French revolution that the records of the National Assembly were to be preserved in the French National Archives and, most importantly, were to be accessible to all citizens (Pannitch, 1996). The question of how to arrange the archival materials in the National Archives, that also drew in Ancien Régime records, was initially addressed through subject classification or the principle of pertinence as a response to catering for the growing number of users, especially historians. Sweeney (2008) suggests that this may have been influenced by library trends at the time which organised their holdings by subject content, who in turn were influenced by the European Enlightenment scientists who used classification systems

for the natural sciences and chemistry. However, by the 1840s, French archivists leaned towards the formalisation of the principle of *respect des fonds,* also known as provenance or the principle of respect for original order, whereby records are organised by source (Brichford, 1989), but could still have subjects applied (Sweeney, 2008). Through the nineteenth century the principle of provenance gradually spread across Europe and was formally ratified at the International Congress of Archivists and Librarians in Brussels, 1910 (Brichford, 1989; Sweeney, 2008).

In modern day, English-speaking countries tend to refer to an archive as a body of records that have been identified and recognised "as having long-term value", and more widely refers "to collections of materials" that are maintained by individuals, organisations, community groups, and governments, or "to the locations where such materials are held" (Yeo, 2017, p. ix). In describing the role of archives, the Constitution of the International Council on Archives (ICA) proposes that they serve to

> constitute the memory of nations and societies, shape their identity, and are a cornerstone of the information society. By providing evidence of human actions and transactions, archives support administration and underlie the rights of individuals, organisations and states. By guaranteeing citizens' rights of access to official information and to knowledge of their history, archives are fundamental to identity, democracy, accountability and good governance (ICA Constitution, 2022)

From this, we see how the function of archives are interlinked with the memory and identity of a society.

Libraries and societies are similarly "interlinked and interdependent" (Ari, 2017, p. 59). Libraries exist for the needs of societies and play a "a vital role" in shaping societies through the transmission and dissemination of "accumulated knowledge through books and other materials" (Ari, 2017, p. 59). White (2012) suggests that libraries serve as societal "gateways to knowledge and culture" while maintaining the balance between authors rights and "safeguarding the wider public interest." Moreover, libraries "are synonymous with education and offer countless learning opportunities that can fuel economic, social and cultural development" across societies (White, 2012). Thus, Padilla (2023) proposes that "Libraries and archives collectively steward global memory - directly supporting education, research, and creativity in small and large communities around the world."

In terms of responsibilities for the collection and preservation of national (tangible) heritage, it is worth noting here that national archives tend to collect and preserve the 'records' and

‘documents’ of government departments and agencies, and may include the collection of records and documents from public bodies, ombudsman agencies or quangos, etc. National libraries tend to collect and preserve the ‘publication’ outputs of a nation state, through a system called legal deposit, which is usually a statutory obligation on publishers to deliver a copy of all new publications. However, libraries may also have special archival collections containing records and documents, while archives may have libraries containing specialised bibliographic collections or rare books. It is widely acknowledged that the collection efforts of national archives and national libraries constitute a major contribution to the preservation of national heritage (Yusuf, 2013; James, 2019; Larivière, 2003; Gooding et al., 2019; Arnold-Stratford & Ovenden, 2020). It is also worthwhile developing an understanding of what constitutes a record, a document, and a publication as these terms are often used interchangeably.

2.1.2 Records, Documents, and Publications

Some schools of thought imply that the term ‘record’ in the physical sense is in some ways self-explanatory. For instance, Cox (2001) suggests that prior to the onset of digital information, most people would have had “a sense about what makes something a record” and offers examples of legal documents, letters, memoranda, receipts, and cheque books as some of the “typical objects” which we considered to be records in both our “personal and professional lives, the result of centuries of organizational and societal activity and evolution” (p. 1). Indeed, Cox (2001) asserts that societies have been “conditioned to these forms by years of personal experience, convention, common sense, and education and training” (p. 1). However, for Cox (2001) the advent of the conception of a “paperless office”, and advances in software technologies which promised “to manage clumps of data” brought about the need for the archival community to seriously reconsider what constitutes a record or document in the same way that electronic publishing has prompted librarians to engage with debates “about the future of the printed book” (Cox, 2001, p. 1). Indeed, the recent decision in Canadian court that the ‘Thumbs Up’ emoji in an email exchange is as valid as a signature on a paper contract and is legally binding illuminates the complexity of what is a record in the digital world (Cecco, 2023).

The International Standards Organisation (ISO) defines a record as “Information created, received and maintained as evidence and as an asset by an organization or person, in pursuit of legal obligations or in the transaction of business” (ISO 15489-1:2016). In differentiating between records and documents, Record Nations proposes that “All records are documents but not all documents are records”, rather, many records may begin as documents and only

become records once they become finalised. Azad (2007) suggests that documents may refer to “a work in progress, which is subject to change” (p. 8), however, documents “can and do become records once they are set in stone, so to speak, and do not undergo changes” (Azad, 2007, pp. 8–9). Hence, a record is not subject to change and consists of “a document or set of documents, all relating to a specific matter that has happened in the past” (Azad, 2007, p. 8). When discussing web-based records, Pennock and Kelly (2006) suggest that the proper management of these records “during the active phase of their life-cycle is vital if authenticity and integrity is to be assured at a later date. In effect, efforts must be expended to ensure Web sites are ‘future-proof’” (p. 1). In addition, while the traditional understanding of a document is “considered to mean text fixed on paper”, the current understanding of a document includes all media/formats inclusive of “photographs, drawings, sound recordings, and videos, as well as word processing files, spreadsheets, web pages, and database reports” (Society of American Archivists, 2005, Document).

Due to developments in computer technology and information communications, the concept of electronic publishing was realised. In the 1960s, electronic publications referred to the use of computers to produce print publications using word processing, typesetting, or mark-up tools. In the 1970s, the first example of an electronic journal was distributed as “a computer readable archival file” and “in the form of computer-output microfiche” (Lancaster, 1995, p. 520). In the 1980s, experiments with internet journals emerged, and email technology allowed for the distribution of e-publications via mailing lists, though this was originally in plain text format (Lancaster, 1995, p. 520; Pettenati, 2002, p. 525), and microfiche also grew as a media form for publication outputs (Schafer, 2020, p. 207). Shifts from “tape to disk storage” during the 1980s also saw the development of processing tools which assisted the organisation of material into databases, seeing the gradual rise of relational database models (Hockey, 2004). Thus, databases (and their datasets) would also need copyright protection as publications, but this non-print type of material challenged the wording provisions of legal deposit, as did other emerging publication technologies. For instance, the development of the CD-ROM in the 1980s offered an effective, low-cost solution for e-publishing and allowed for good quality graphics and images (Pettenati, 2001, p. 525). CD-ROM began to be replaced as a medium from the end of the 1990s, due to the development of the web, and the increasing availability of the internet (Waniata, 2018). Electronic publishing also brought with it a new dimension for legal deposit schemes, which up to this point had been mostly print-centric (Muir, 2005, p. 4).

Today, an electronic or digital publication can refer to a text encoded version of a print publication (e.g., encoded using XML/TEI), a scanned/digitised version of a print publication

(e.g., a PDF or microfiche), or a born digital publication where there is no print parallel. Examples here include web pages and blogs on the web; e-zines and e-newsletters distributed via email technologies; information disseminated on bulletin board systems (BBS) via the internet; social media delivered through web or mobile phone technologies; and even interactive CD-ROMs (e.g., CD-ROM encyclopaedias). It also includes other types of publications that can be hosted on the web such as podcasts, videos, digital scholarly editions, interactive databases containing bibliographies, statistics, spatial data etc. (Boston, 1998; Taylor, 2013).

In offering a distinction between a document and a publication, the United Nations Regulations for the Control and Limitation of Documentation (ST/AI/189/Add.3/Rev.2) defines a document as a "text submitted to a principal organ or a subsidiary organ of the United Nations for consideration by it, usually in connection with item(s) on its agenda" (United Nations Secretariat, 1985), whereby examples include documents which are "issued for or under the authority of intergovernmental bodies under a United Nations document symbol and include all official records and meeting records of organs or conferences of the United Nations" (The United Nations Office at Geneva, n.d.). On the other hand, it defines a publication to be "any written material which is issued by the United Nations to the general public", with examples including "major studies and reports, monographs, edited volumes, statistical compilations, conference proceedings, journals, serial publications such as yearbooks, the United Nations Treaty Series and other international law publications" (United Nations Office at Geneva, n.d.). The United Nations Office at Geneva further adds that publications include "print or electronic form, including as mobile applications, and in any other format or media as technology evolves" (n.d.). Thus, when it comes to the website of the United Nations, it contains and provides access to the records, documents, and publications of the United Nations.

2.2 Societies, Communications, and Culture

The sociologist Anthony Giddens (1997) describes the concept of a society as "a group of people who live in a particular territory, are subject to a common system of political authority, and are aware of having a distinct identity from other groups around them" (p. 585). Elsewhere, Conerly et al. (2021) describe a society as "a group of people who live in a defined geographic area, who interact with one another, and who share a common culture" (p. 8). Wikipedia (2002+) offers a description of a society as "a large social group sharing the same spatial or social territory, typically subject to the same political authority and dominant cultural expectations." This next section examines the nature of a society as a large social

group who interact and communicate, have multiple things in common (e.g., territory, language, traditions, culture, political institutions), and have some awareness that they differ from other social groups. Therefore, it is worth examining some of the underlying reasons for the formation of social groups in the first instance.

2.2.1 Social Groups and Identities

As humans, we need to interpret the world around us by filling it with meaning. Meaning is appropriated through classification and categorisation systems which is an unavoidable process and a necessary condition for the human mind (Allport, 1959, p. 20; Bodenhausen et al., 2012, pp. 318–319; Jenkins, 2008b, p. 105). How we define ourselves and others, and how others define themselves and others (including us), is similarly dependent on a system of classification, namely social categorisation. This process adds meaning to identity which would otherwise be ambiguous. Understanding the concept of identity in academic frameworks can be confusing (Ashmore et al., 2004; Brubaker & Cooper, 2000).[2] Richard Jenkins (2008b) offers a sociological framework on social identity and argues that "all human identities are, by definition, social identities" (p. 5; p. 17). On the other hand, from a psychological level, Ashmore et al. (2004) refer to social identity as a collective identity, which they differentiate from personal identity, whereby "collective identity is explicitly connected to a group of people outside the self, personal identity typically refers to characteristics of the self that one believes, in isolation or combination, to be unique to the self" (p. 82). While Jenkins (2008b) acknowledges that for some theorists "*collective* identity and *individual* identity are typically understood as different kinds of phenomena", Jenkins maintains that this differentiation is somewhat rejected by social scientists as "individual and collective identifications only come into being within interaction [and] each emerges out of the interplay of similarity and difference" (pp. 37–38). So, from Jenkins's analysis, it could be surmised that all human identities are social identities that arise from processes of identification through interaction and the interplay of sameness and difference.

Social psychologists such as Turner et al. (1994) sum up social identity as "the social categorisations of self and others" (p. 454). Social categorisation is a process by which

[2] Brubaker & Cooper (2000) suggest that as the term identity permeated across different fields of discipline its usage acquired flexible definitions, and so, they propose that the term is no longer useful. In agreement, a study of identity literature by Ashmore et al. (2004) found similar ambiguity and "conceptual confusion", but, unlike Brubaker & Cooper (2000), they believe that "the concept needs to be better articulated rather than abandoned or severely restricted" (Ashmore et al., 2004, p. 82).

human beings are classified into a range of social categories such as men, women, artist, labourer, Catholic, Protestant, Republican, Democrat, French, Spanish, and so on (Bodenhausen et al., 2012; Strangor, 2004). Therefore, social identification is the process by which people perceive their self-categorisations to have similarities or dissimilarities with other social categorisations (Ashforth & Mael, 1989). A shared social identity may be based on common characteristics which are determined or ascribed, such as race, ethnicity, language, age, and gender, or based on "achieved states such as occupation or political party" (Ashmore et al. 2004, p. 81). Furthermore, sharing a social identity does not necessitate that categorical members need to know or be in contact with other members of the same category (Deaux, 2001; Stangor, 2004). Jenkins (2008b) discusses how (social) identification involves a practical process which is ongoing, as the dynamics of (social) identity is "never a final or settled matter", nor is one's (social) identity confined to the singular, rather, it is "multi-dimensional" (Jenkins, 2008b, pp. 16–17). Multiple social identities enable people "to adopt various roles and adapt to a variety of social contexts" (Code & Zaparyniuk, 2009, p. 92). So, if identity is multi-dimensional, and it is not inherent or fixed, this implies that social identities are "somewhat negotiable and flexible" (Jenkins, 2008b, p. 18). So, for Jenkins (2008b), people share aspects of their (social) identity with some, but not with others, and this may be open to change or renegotiation.

In developing an understanding of social groups, Jenkins (2008b) asserts that making sense of who is who revolves around the identification of what people have in common with a social categorisation or social group, therefore, this also coincides with "the recognition of other groups or categories from whom they differ" (p. 23). However, Jenkins (2008a; 2008b) infers that there is a difference between a social category and a social group. Jenkins (2008b) proposes that it is the process of "internal collective definition" which allows for a social group to come into existence, therefore it is a process of internal identification by its members and "the relationships between them" (p. 106). On the other hand, social categories are externally "identified, defined and delineated by others" (Jenkins, 2008a: p. 56; p. 83). This suggests that, by necessity, the social construction of a group identity coincides with social categorisation and internal group identification as a dual social process. Moreover, 'we' as a group may define ourselves in terms of shared characteristics of a social category, but, 'we' (as a target) may be externally defined by others (as perceivers) to belong to an alternate social category, even though 'we' as a social group might not accept or recognise that 'we' belong to that category (Bodenhausen et al., 2012, p. 319; Jenkins, 2008b, p. 105; Jenkins, 2008a, p. 57). This further correlates with the behavioural processes of ingroups and outgroups.

In *Folkways* (1906), the sociologist William Sumner was one of the first to describe the distinction between in-groups and out-groups as a phenomenon of human behaviour. In explaining this concept, Bodenhausen et al. (2012) suggest that when a person places another person in a social category, they are "likely to consider [their] own status with respect to that category", that is as a categorical member or non-member (p. 319).[3] This allows for (individual) people to evaluate whether they have a (psychological) connection or sense of belonging to this social category, and if so, this becomes an in-group—and if not, this then becomes an out-group (Bodenhausen et al., 2012, p. 319; Stangor, 2004, p. 113). In explaining "intergroup conflict", the social sciences often identify how individuals are prone to categorise themselves in terms of in-groups (us) and out-groups (them) (Schmid et al., 2010, p. 457). This is further supported by the minimal group experiments conducted by social psychologists Tajfel and associates (e.g., Tajfel 1970; Tajfel 1971; Tajfel et al., 1974) which aimed to develop a deeper understanding of intergroup discrimination through the minimal conditions in which discrimination would occur. The findings of Tajfel and associates imply that even "the mere perception of belonging to two distinct groups" (i.e., social categories) is enough to produce elements of in-group favouritism and discrimination towards out-groups (Coenders et al., 2004, p. 8). However, this does not mean that in-groups and out-groups will resort to conflict, rather it accentuates differences, which in turn may be attributable to causes of conflict. It further highlights how social groups can exercise power through the processes of inclusion and exclusion as a product of in-group and out-group behaviour.

It should also be noted that within any given society there exist individuals who identify as belonging to ethnic groups or communities. For some commentators, to be a member of an ethnic group implies "shared origins" (Senior & Bhopal, 1994, p. 327), and/or is defined as a collective (social) identity in terms of shared culture, language, or religious traditions (Deaux, 2001, p. 4). Further debate on ethnicity evolves around whether it is "primordially given or optionally cultivated" (Gleason, 1983, p. 919). Primordialists infer that ethnicity is a "basic element in one's personal identity that is simply there and can not be changed" (Gleason, 1983, p. 919). Others consider ethnicity through the ideology of constructivism which explains "ethnic identities as products of social constructions, human actions and choices" (Blizzard, 2006, p. 3). For example, cultural optionalists consider that ethnicity is not a stamp

[3] Strangor (2004: 113) considers this as self-categorisation, in that "a person thinks about himself or herself (rather than thinking about another person)" (p. 113).

"impressed on the psyche", rather, "ethnicity can, within certain limits, be assumed or put aside by conscious choice" (Gleason, 1983, p. 919).

In defining an ethnic group, Smith (1993) suggests it is a "named human population with a myth of common ancestry, shared memories and cultural elements, a link with an historic territory or homeland and a measure of solidarity" (pp. 29–30). On the other hand, grounded in social anthropology, Barth (1969) suggests that "ethnic groups are categories of ascription and identification by the actors themselves" (p. 10). Barth (1969) was more interested in the "ethnic boundary that defines the group, not the cultural stuff that it encloses" (p. 15). Rather, for Barth, "the essence of an ethnic identity is to emphasise the boundary between insiders and outsiders" (Nic Craith, 2002, p. 137). Michael Banton (1983) presents a useful explanation that differentiates between race and ethnicity, in that "'race' is a categorical identification denoting 'them', based on physical or phenotypical characteristics, while ethnicity is the cultural group identification of 'us'" (cited in Jenkins, 2008a, p.50). Therefore, for Banton (1983) ethnicity is "voluntarily embraced", while "racial identifications are imposed", however Banton maintains that both are social constructions "albeit perhaps with different force" (cited in Jenkins, 2008a, p. 51).

Community also has its fair share of definitions and theoretical implications (Cohen, 1985; Delanty, 2009). Cohen (1985) offers a "reasonable interpretation" on community as: "the members of a group of people [who] (a) have something in common with each other, which (b) distinguishes them in a significant way from the members of other putative groups" (p. 12). Therefore, Cohen (1985) suggests that "the boundary marks the beginning and end of a community", but stresses that "not all boundaries, and not all the components of any boundary, are so objectively apparent" (p. 12). Rather he proposes that they might be thought of "as existing in the minds of their beholders" and refers to this as the "symbolic boundary" between communities (Cohen, 1985, p. 12). Symbols by themselves are meaningless, however, when they are given meaning, they are "multi-vocal", that is they "do not communicate a single proposition, but rather a collection of propositions, ideas and emotions" (Bryan & Gillespie, 2005, p. 13). So, symbols may be interpreted differently by actors or agents inside or outside the community. While Cohen's (1985) description of community might easily be compared to Barth's (1969) description of an ethnic group, however, Ruane and Todd (2004) suggest that "Ethnic categories can exist without communities, [. . .] and strong and intense communal bonding infused with a sense of kinship may also exist relatively detached from ethnicity" (p. 12). Nonetheless, both Cohen's and Barth's interpretations suggest that it is the boundary which delineates 'us' from 'them'.

Nationality is also perceived as an identity but may be separated analytically into two distinct parts. Legal nationality is ascribed through the citizenship regime of a state and so this implies nationality as citizenship by way of "identification" (Hayward & Howard, 2002). According to Bellamy (2008), "To be a citizen is to belong to a given political community" (p. 52). Since sovereign states are political communities with boundaries, Walzer (1983) suggests that the state has "the right to exclude or include whomever they choose" as the state retains sovereignty to formalise the processes and determine the criteria for granting citizenship (cited in Tracy, 2000, p. 10). Pierson (1996) describes a modern state as comprised of elements such as a bounded territory, sovereignty, authority, governance, citizenship, and legitimacy for the use of violence. Indeed, for Weber (1978), "The claim of the modern state to monopolise the use of force is as essential to it as its character of compulsory jurisdiction and continuous operation" (p. 56). This does not necessitate that a state is solely organised around the use of force to administer governance. In the normative sense, the will of the state's authority also depends on the consent of its citizens to be governed. Thus, in broad terms, the state uses two strategies for ensuring compliance with its rules: the engineering of consent and the legitimate use of force (Webber, 1978, pp. 55–56).

On the other side of this, nationality may also be a "self-definition" in terms of an ethno-cultural identity and thus, this implies that nationality is a shared identity by the members of a nation (Hayward & Howard, 2002). Whilst defining a nation for the most part is problematic, Smith (2004) suggests that it is the ethnic group majority which "provide the unifying elements (in terms of land, language, law and customs) of the modern nation" (cited in Kornprobst, 2005, p. 405). However, Benedict Anderson (2006) proposes that a nation is "an imagined political community and imagined as both inherently limited and sovereign" (p. 6). Thus, for Anderson (2006) a nation is imagined by people who perceive that they belong there, however, they may never know, meet, or speak with other members, "yet in the minds of each lives the image of their communion" (p. 6). Anderson (2006) asserts that the protective and emotive influences experienced by those who imagine their unity as members of a nation cannot be dismissed and, for him, it is from this that the ideology of nationalism grows. In differentiating between the nation and a nation state, Mukherji (2010) posits that recurrent discourses provide for the nation to be a "cultural/ethnic category" while a nation state is perceived to be "a specific form of state, which exists to provide a sovereign territory for a particular nation, and which derives its legitimacy from that function" (p. 2). However, while legal nationality as citizenship implies membership of the nation state, it does not guarantee membership of the nation of the state. This is apparent

in the case of Irish Travellers who as Irish citizens were customarily left outside the boundaries of the Irish nation and the social construction of Irishness (Lentin, 1998). Moreover, others have argued that the rise in immigration from the Millennium has led to a renegotiation of citizenship and what it means to be a member of the Irish nation state (Honohan, 2007; Fanning & Mutwarasibo, 2007; Collins, 2010; Ní Mhurchú, 2011).

Following the Irish War of Independence (1919-1921), the newly established Irish Free State (1922) came close to being an ideal-type nation state as the state boundaries were drawn to serve a 93% Catholic majority (Hug, 2001, p. 25). However, following the Irish Civil War (1922-1923) the social fabric of Ireland was divided, and the first undertaking of the Irish Free State was "to solve the problem of integration and solidarity" through the connection of "nation and community" (O'Carroll, 2002 cited in Geoghegan, 2008, p. 129). This was evident in state politics for a few decades which endorsed notions of commonality through the Irish language and Catholicism. Indeed, the development of the new state had an emphasis on an "Irish-Ireland" which became increasingly "protectionist and isolationist" (Fanning, 2010, pp. 399–400). This persisted until the 1950s when economic developmentalists began to move from protectionism to embrace "economic and human capital reproduction as utilitarian nation-building goals" (Fanning, 2010, pp. 399–400). This was pursued through promoting education to create a skilled labour force, the development of more liberal trading agreements, and then the entry of Ireland as a member to the European Economic Community in 1973 (Fanning, 2010, pp. 399–406). Nonetheless, the Republic of Ireland still remained as an "Irish-Ireland" as it was not predisposed to noteworthy levels of immigration or a momentous increase in the demand for citizenship until the period of economic growth known as the Celtic Tiger era from the mid-1990s until the global recession of 2008. Macrotrends offers a good visualisation of the growth rate of immigration from the 1980s to 2015, vis-à-vis the percentage of the population (Figure 2.1).

With the start of the Celtic Tiger boom, Irish labour market demands surpassed expectations, thus large-scale immigration was seen as a means for supplying the demand for continuing economic success. Initially, the government targeted Irish skilled workers abroad to return home, a strategy that was pursued by means of playing on ethnic ties and a patriotic duty (Hayward & Howard, 2007). Thus, the scheme for enticing non-national economic migrants coincided with the decline of a strategy aimed at enticing the return of Irish-born skilled workers. However, the 'new Irish' who contributed to the economic growth of the country during the Celtic Tiger, some of whom then became citizens, were merely offered a "woolly notion" of interculturalism (Munck, 2011, p. 4). Others have argued that this generation of new Irish "will grow up as strangers in their own country, forever seen as an alien

contaminant within the true blood of the nation-state" (Maguire & Cassidy, 2009, p. 18). Again, this demonstrates how legal nationality as citizenship does not ensure affiliation to the nation of the state. Moreover, one should also consider that since the conception of the Irish Free State up until the 2000s, the social, cultural, and political institutions of the Republic of Ireland have been accustomed to catering for a nation of settled white Irish Catholics (Howard, 2016, p. 169). Thus, it could be argued that this will have an influence on the production and preservation of Irish national heritage. Therefore, there will always be a need to consider how collection development policies for national heritage collections may revolve around a dominant hegemonic social group, at the cost of excluding representations from ethnic minorities, societal sub-groups, or alternative communities.

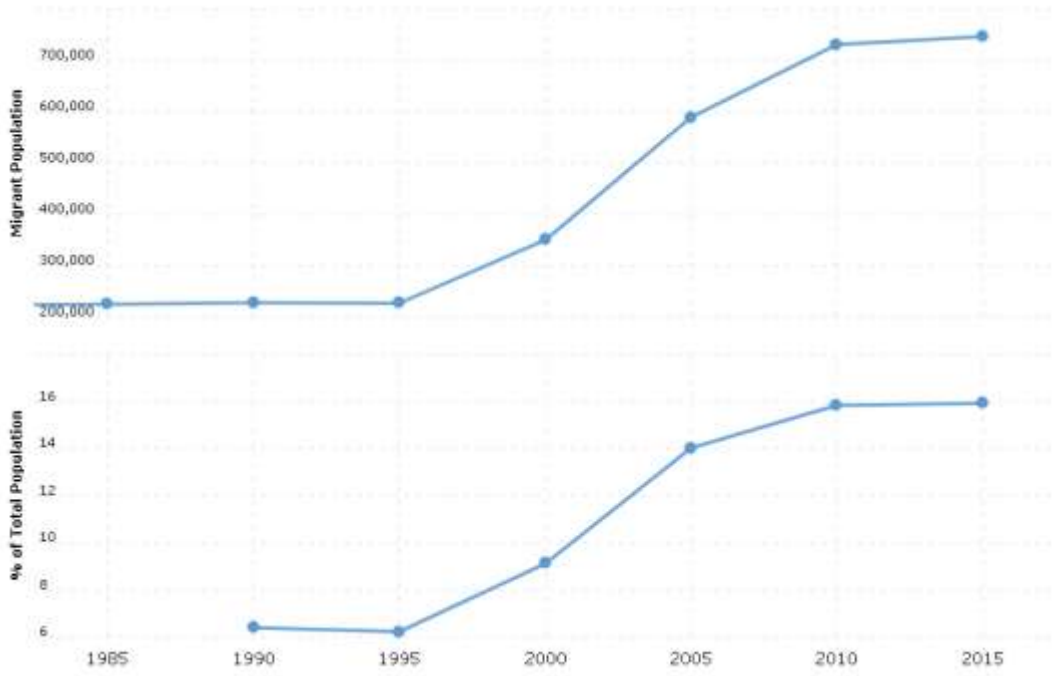

Figure 2.1: Growth rate of immigration from the 1980s to 2015, vis-à-vis the percentage of the population (Macrotrends)

2.2.2 Communications and Culture

Conerly et al. (2021) advocates that "consciously and subconsciously" humans are constantly striving to make sense of their surrounding environments, their world, and use symbols as a form of communication (e.g., gestures, signs, objects, words, signals, etc.). These symbols provide cues to convey ideas, common understandings, or shared experiences. For example, language is a "symbolic system through which people communicate and through which culture is transmitted" (Conerly et al., 2021, p. 77).

Kroeber and Kluckhohn (1952) also offer this connection through their description of culture which encompasses,

> patterns, explicit and implicit, of and for behaviour, acquired and transmitted by symbols, constituting the distinctive achievement of human groups, including their embodiments in artifacts; the essential core of culture consists of traditional (i.e., historically derived and selected) ideas and especially their attached values" (Kroeber & Kluckhohn, 1952, p. 181).

It can also be argued that it is within communications that societies are created and sustained, and culture is transmitted (Cooley, 1909, Dewey, 1916; Kroeber & Kluckhohn, 1952, Conerly et al., 2021). Dewey (1916) points out how there is "more than a verbal tie between the words common, community, and communication", it reflects that individuals live in communities "in virtue of the things which they have in common; and communication is the way in which they come to possess things in common" (p. 5). Moreover, communications and media "propagate the values and schemas of a culture through the repeated interaction and exchange enabled by the communications process" (Media Texthack Team, 2014). This repeated interaction allows for culture to be fostered in public consciousness and memory.

While definitions and theories of communications vary, it might be useful to start off with a basic understanding of communications as "the process of generating meaning by sending and receiving verbal and nonverbal symbols and signs that are influenced by multiple contexts" (University of Minnesota Libraries, 2016, p. 3). The cultural theorist Raymond Williams (1973) indicates that in English, the oldest meaning for the term communications can be summarised "as the passing of ideas, information, and attitudes from person to person" (p. 17). Although definitions of the term communications vary. Williams (1973) points out that since the onset of the Industrial Revolution, the term communications was also extended "to mean a line or channel from place to place", and in this context refers to "ways of travelling and carrying" (Williams, 1973, p. 17). However, Williams (1973) notes that this has since become confusing due to developments in the twentieth century of telegraphy, photography, radio, film, television, and the computer, and suggests that the term communication is better suited to these types of developments, while transport is a better term to use when referring to the "physical means of travelling or carrying" (p. 17). Nonetheless, as Williams (1973) points out, both meanings for the term communications "will go on being used" (p. 17; c.f. Behringer, 2016).

In differentiating between communications and media, it is useful to consider media as the plural form of medium which represents some type of channel of communication (Rousse, n.d.). Media are made up of various 'mediums', with various dimensions (e.g., cave art, clay tablet, tomb inscription, art canvas, print document, film, radio, television) which can encompass multiple types or genres. For instance, *print media* genres might refer to newspapers, magazines, comics, or pamphlets. In the field of information studies/sciences, media may also refer to *storage media* (e.g., *fixed media*: hard drives; *reusable media*: floppy disks, CD-ROMS, USB memory sticks, etc.); or as *transmission media* (e.g., videos, sound recordings, podcasts, etc.) (Rousse, n.d.; Society of American Archivists Dictionary, 2005, Media). Brügger (2019; 2018; 2016) examines how various types of digital media (materials) come with their own form of digitality, "that is, a specific way of being digital" (2018, p. 5). In this respect, Brügger (2019; 2018; 2016) puts forward three major categories for distinguishing different types of digital media. Brügger (2019; 2018; 2016) offers the typologies of: digitised media, as media that is originally non digital (such as analogue materials that have been digitised, e.g., paper or parchment documents, print newspapers, print photographs, negative films, or any "form of transformation of analogue material into digital form" (Brügger, 2016, para 28). Then there is born digital media, which is 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 (Brügger, 2019; 2018; 2016).

It is Marshall McLuhan who provides us with a theory of communications which further demonstrates how societies, communications and culture are inextricably linked, although his ideas are more widely received today than when they were first presented back in the 1960s and 1970s. McLuhan examined communications through the lens of media and transformation theory in terms of "what these media do to the people who use them" (McLuhan, 1970). For example, what did the invention of "writing do to the people who invented it and used it?" (McLuhan, 1970). Indeed, McLuhan is more interested in the way that societies have been shaped "by the nature of the media by which men communicate than by the content of the communication" and uses the alphabet as an example of

> a technology that is absorbed by the very young child in a completely unconscious manner, by osmosis so to speak. Words and the meaning of words predispose the child to think and act automatically in certain ways (1967).

McLuhan was particularly interested in developments in communications technologies in the era of the radio, film movies, and television, labelling these events as the electronic

period, after the print period. When discussing these advances, McLuhan discusses how these developments are

> reshaping and restructuring patterns of social interdependence and every aspect of our personal life [...] forcing us to reconsider and re-evaluate practically every thought, every action, and every institution formerly taken for granted. Everything is changing –you, your family, your neighbourhood, your education, your job, your government, your relation to 'the others.' And they're changing dramatically (1967).

It is within this context that McLuhan discusses his theories on the "medium as the message" whereby it "is impossible to understand social and cultural changes without a knowledge of the workings of media" (1967).

When discussing the development of television as a medium, McLuhan proposes that any new form of media,

> comes into the foreground of things, we naturally look at it through the old stereos. We can't help that. This is normal, and we're still trying to see how our previous forms of political and educational patterns will persist under television. We're just trying to fit the old things into the new form, instead of asking what the new form is going to do to all the assumptions we had before (1960).

On the plus side electronic communications technologies brought about benefits of entertainment, education, multicultural and intercultural experiences, interconnecting communities, societies, and diaspora (McLuhan, 1967). On the downside, McLuhan points out how "Innumerable confusions and a profound feeling of despair invariably emerge in periods of great technological and cultural transitions" whereby anxieties are heightened for the most part as a "result of trying to do today's job with yesterday's tools-with yesterday's concepts" (1967). These words resonate even more so today, particularly for archivists and librarians, due to the rapid advances with the internet, web, and social media technologies over the last three decades alone. As pointed out by Muldoon (1997), shifts in information communications technologies have been marred with a similar phenomenon, through "the loss of information that occurs when a new technology is introduced and supersedes an older one" (p. 50). Lyman (2002) notes how the focus is on the creation of new information and not the preservation of the old. Moreover, while training and education is directed towards the latest communications technologies, it negates the need for training and education in the long-term preservation of information created by these new technologies (Lukow, 2000; Byrne et al. 2024 forthcoming). Thus, to some degree, archivists and librarians have always faced new and evolving challenges due to

transformations in media and communications technologies. The histories of the preservation of film and television heritage offers good examples of this (McKernan, n.d., Ballhausen, 2011; Giuliani & Negri, 2011; Ide et al., 2002; Wright, 2017). Indeed, it could be argued that the work and experiences of film, photography, and television archives of the twentieth century set the foundations for the evolution of born digital archiving.

For Castells (2007), web and social media technologies also brought changing patterns and dynamics of power-relations in the communications landscape due to the evolving nature of the network society. According to Castells (2007, 2012), we now live in a "network society" where collective identities and online communities' transverse beyond borders, nation states and continents, and where a horizontal communications framework of "mass self-communication" is now utilised to bypass the vertical framework paradigm of the mass media gatekeepers. It is in the space of mass self-communication that the new era of social movements is formed (e.g., Occupy, Arab Spring etc.), and it is this autonomous space which allows for new social movements to generate attention, garner support, and orchestrate mobilisations for collective action (Castells, 2012). Castells (2007) suggests that in order to understand how these transformations are being brought about, there is a need to examine them "in a social context characterized by several major trends", such as mass communication and media politics, scandal politics, and the crisis of political legitimacy (p. 239). For example, Castells (2007) illustrates the benefits of horizontal communication networks for political actors to gain votes, to sabotage their opponents and also for the public to become more engaged in political debate (pp. 254–256).

Castells' (2007, 2012) views on media as a space of power-making has validity, but this is nothing new. The development of electronic communications and media technologies throughout much of the twentieth century have to some degree presented an effective landscape for discursive appropriation in the politics of representation, and the battle to establish what Stuart Hall identifies as "preferred" meanings (Hall, 1980, p. 57). The politics of representation is concerned with the "competition over the meaning of ambiguous events, people, and objects in the world" (Meehan, 1996, p. 241). This competition takes place through discourse and media, as individuals and groups attempt to establish their view as the proper or "preferred representation" (Wenden, 2005, p. 90: Meehan, 1996, p. 241). Advocates of varying views use different "strategies to ensure that their framing of the nature of a particular issue predominates" (Wenden, 2005, p. 91). If a strategy is successful "a hierarchy is formed, in which one mode of representing the world [. . .] gains primacy over others" (Meehan, 1996, p. 241). This hierarchy then becomes a "dominant-hegemonic" in producing and reproducing representations associated with "preferred meanings" (Hall,

1980, pp. 57-59) which quite often “saturate commonsense” (Wenden, 2005, p. 91). Therefore, if representations add meaning to ambiguity and the politics of representation is the continuing struggle for control to produce representations which establish ‘preferred meanings’, this also brings into question how preferred meanings are represented, renegotiated, and re-presented through the organisation, interpretation, and decision-making of heritage organisations; and how this impacts the transcendence (or not) of preferred meanings into commemorative memory for the next generation.

One should also consider how electronic communications and media technologies have been used as tools to manipulate and control societies and social groups, whether this is due to interfering with the availability of the internet during the Egyptian ‘Arab Spring’ (Olukotun & Micek, 2016), or the use of the radio to spread messages of hate and inciting ethnic tensions in the lead up of events which led to Rwandan Genocide in 1994 (Kellow & Steeves, 1998). Also, in the latter part of the twentieth century, concerns were raised about increased concentrations of mass media ownership of newspapers, magazines, television, motion pictures, and books (Gamson et al., 1992, pp. 375–377), and how the phenomenon of mass media monopolies influenced the politics of representation (Herman & Chomsky, 1988; Gamson et al. 1992; Kamalipour, 1997; McChesney & Schiller, 2003). For example, Gamson et al. (1992) discuss how “media-generated images of the world” are used “to construct meaning about political and social issues”, however, the

> lens through which we receive these images is not neutral but evinces the power and point of view of the political and economic elites who operate and focus it. And the special genius of this system is to make the whole process seem so normal and natural that the very art of social construction is invisible (p. 374).

McChesney and Schiller (2003) highlight how, for much of the twentieth century, corporate media and governments could be seen as partners in establishing preferred meanings as they had control of the spheres of communication which reached the masses.

On the other side of this, the advent of the internet, web, and social media generated spaces for alternative discourses that foster a greater pluralism of perspectives, and by-pass traditional modes of media gatekeeping (Castells, 2007; 2012). Moreover, it has driven the growth of online communities with their own social and cultural values and therefore, also their own heritage, which is inclusive of an era of ‘influencers’ and ‘cancel culture’. Social media influencers are individuals who tend to have a large follower base and “make money by posting regularly and engaging their audience” with platforms such as TikTok, YouTube, Twitter, and Instagram (Lewis, 2023). However, influencers are also subject to a process

known as “cancelling” whereby their followers stop supporting them “because of something they posted and did that is not socially acceptable in our society” (Lewis, 2023). Hence, when it comes to the preservation of national digital heritage on the web and social media, there needs to be ongoing discussion on what gets captured and why, and how it reflects the ongoing transformations in societies, communications, and culture. There is also a need to consider what constitutes national heritage in line with the politics of representation, preferred meanings, and alternative discourses, and how these concepts influence what is included or excluded in the preservation of national heritage, and/or how this translates into the gaps and silences which are inherent in national archives and libraries.

2.2.3 Gaps and Silences

Over the past decades there has been much discussion on the gaps and silences in archives and libraries, and how such gaps have come into being. Carter (2006) attributes silences in part, to “the manifestation of the actions of the powerful in denying the marginal access to archives” which has “a significant impact on the ability of the marginal groups to form social memory and history” (p. 215). For Fowler (2017) archives and the sources they contain are “neither natural or neutral”, rather they are “created” by a human process from the production and selection of records/documents to the cataloguing and delivery of the same (p. 1). Fowler (2017) posits this as a main reason for “so many silences” (p. 2). Fowler (2017) further describes how those in power may purposefully thwart the creation, preservation, and access of records, and silences in the archive can result from a culture of secrecy, or wilful or convenient destruction. Therefore, this will also have an impact on the politics of representation, and the production and preservation of heritage. However, absences in archives and libraries are not a recent phenomenon.

Yeo (2017) demonstrates how absences in the archive (or libraries) were chronicled by Thomas Walsingham who lived during the fourteenth-fifteenth century, and documents how a peasant’s revolt in 1381 led to demands to see a charter for civil liberties of the peasants which was supposedly stored in a monastery. However, the charter could not be found in the monastic archive. Thus, the claims by the rebels “stayed a matter of conjecture and contestation”, and then with a lapse of time there was no way of proving whether a document of “ancient liberties” actually existed or was “merely a fable” (Yeo, 2017, p. x). Elsewhere, Harley (1988) discusses cartographic silences through a political reading of maps from the early modern period in Europe, starting from the sixteenth century. Harley (1988) suggests that cartographic silences “are contributed by many agents in the mapmaking process, through the stages of data gathering to those of compilation, editing, drafting,

printing, and publication" (p. 57). Therefore, when it comes to assessing cartographic silences, one needs to "be aware not only of the geographical limits to knowledge but also of the technological constraints to representation, and of the silences in the historical record owing to the destruction of evidence" (p. 57). Harley (1988) maintains "that which is absent from maps is as much a proper field for enquiry as that which is present" (p. 58).

Silences in the heritage of archives and libraries may also be due to displacement or loss as a consequence of conflict, war, decolonisation, custodial neglect, arson, theft, and natural disasters (Inkster, 1983; Bastian, 2001; Tough, 2009; Banton, 2012; Onyeneke, 2017; Kuzucuoglu, 2014). Archives and libraries have been lost through deliberate destruction, as happened in Nazi-era Germany (Ovendun, 2020), or by accident, as happened in the Florence floods of 1966 (Horne, 2016). However, because of the existence of catalogues, inventory lists and registries there is often a good idea of what was lost and can even perhaps plan for some type of reconstruction through the sourcing of duplicates elsewhere. Such is the case with the Virtual Record Treasury of Ireland (https://virtualtreasury.ie/) who are attempting to simulate a working construction of the past records of the Record Treasury of the Public Record Office of Ireland (PROI) through the sourcing and digitisation of duplicate records which were destroyed during the opening weeks of the Irish Civil War in 1922 (Wood, 1930, p. 35). This will be discussed further in Chapter 5. The challenge of the digital is that once lost it may be irretrievable. Moreover, in the digital age, silences in the digital heritage of archives and libraries exist due to legal, ethical, curatorial, financial, technical, temporal, social, and political circumstances and will be discussed in more detail in the next sections.

2.3 Preservation of Digital Heritage

The advent of computational records, electronic information, multimedia and born digital materials have all presented concerns for archivists and information professionals regarding the appraisal, storage, and long-term preservation of such heritage. In the early 1970s, Fishbein speculated that unless archivists were brought up to speed to deal with the challenges of appraising and preserving computational records, "about one million reels of tape in the Federal Government and more elsewhere will be erased without any archival judgments on the continuing value of the information they store" (Fishbein, 1972, p. 35). In 1985, the Committee on the Records of Government (1985) cautioned that the "United States is in danger of losing its memory" due to the shift from paper to electronic records and the instability of maintaining electronic materials (p. 9). The 1996 report, Preserving Digital Information by the Commission on Preservation and Access and the Research

Libraries Group (RLG) further identified concerns for the preservation of electronic and multimedia materials stressing "the need to protect against both media deterioration and technological obsolescence" (Waters & Garrett, 1996, p. iii, p. 5). For example, media that was stored on nitrate film and magnetic tapes often deteriorated beyond redemption; and information stored on older floppy disk versions were at risk of being unreadable by upgraded technology (Besser, 2000; Lyman, 2002). Terry Kuny (1997) coined the term "Digital Dark Age" to highlight the loss of historical information due to outdated file formats, the upgrading or obsolescence of software and hardware, and the loss of information on the internet.

2.3.1 Born Digital Heritage

Ide et al. (2002) discuss the challenges for archiving digital television, due to its characteristics of being "a hodgepodge of media types and formats" (p. 67). They propose that, in many ways,

> the dilemma of archiving digital content is the same as it was for analog: how do we preserve the substance of a medium while its physical containers decay or grow obsolete? For analog products, standard practice recommends procuring appropriate shelf space within a controlled environment. Digital objects may be handled in similar fashion-that is, as shelved artifacts-but this approach avoids examining the qualities that make digital both attractive and perilous for productions (pp. 67–68).

Reference to the preservation of electronic mail (email) was highlighted by the lawsuit *Armstrong v. Executive of the President* (1989). The lawsuit was first filed to prevent the deletion of emails created by the Reagan and Bush White House administration. Consequently, the case set a precedent for the formal acknowledgement that emails formed a part of the Presidential records to be handed over at the end of term, for appraisal and preservation by the state archivist (Bearman, 1993). Indeed, referring to the Republic of Ireland in 1997, Michael Cunningham of *The Irish Times*, also discusses the preservation of emails, and asks:

> If a digital national archive is important for the historians of the future, where is Ireland's digital archive? Which national agency in Ireland should - or could - be responsible for saving and preserving today's email and other electronic objects? (Cunningham, 1997b, p. 18).

While Cunningham notes that "most employees in central and local government" did not appear to have access to or use email technology and points out that the "problem might

seem far off", he suggests that "the longer the State postpones decisions in such areas, the bigger the chunk of our country's digital history that future generations will lose forever" (Cunningham, 1997b, p. 18).

Additionally, in the Republic of Ireland, from 1997 up to 2012, the Director of the National Archives of Ireland (NAI) repeatedly advised on the "pressing need" for the long-term preservation of electronic government records as outlined below.

> In the annual reports since 1997, attention has repeatedly been drawn to the pressing need for action to ensure the long-term preservation of records in electronic form. Much of the business of Government is now transacted electronically and it is essential that a legal and regulatory framework, and resources and systems be put in place to ensure that the electronic records generated can be managed and preserved into the future, thereby facilitating Government accountability and preservation of the national memory (NAI, *Report of the Director for 2012*, p. 17).

Also, in an article in *The Irish Times* in 2012, the keeper of the NAI, Tom Quinlan highlighted how the Irish Government still did not "have a designated system for preserving and retaining" email (Quinlan cited in Fagan, 2012).

Regrettably, over twenty years since the problem was identified, the Irish government has still not come to terms with the preservation of electronic records, nor does it seem to have a formal policy for record keeping in any electronic format. This is pointed out, year after year, by the reports of the Director of the NAI from 2014 up until 2020. These reports identify risks related to the lack of a "comprehensive formal records, management policy for State" and the "Loss of electronic records and archives or access to them, due to degeneration of storage media and/or redundancy of operating systems" (NAI, Reports of the Director for 2014-2020).

On a more positive note, the NAI produced an ambitious strategy in 2021 to deal with the information age, which includes "a digital transformation programme [and] a new framework for records management across government" (NAI, n.d., News; NAI, 2021b). Of course, achieving the goals of the strategy will depend on "improved funding, an enhanced infrastructure and [...] improved staffing resources" (NAI, 2021b, p. 6). It will also depend on the universal adoption of a framework for governmental records management, by civil servants, local government, and the Oireachtas, and this will require an organisational cultural shift which may be more difficult to negotiate (Denning, 2011).

From the mid-1990s, discussions on the preservation of electronic information would further coincide with concerns about the vulnerability of information and documents on the internet and the web. In the spring of 1996, the lead technology officer at Microsoft, Nathan Myhrvold, started an email conversation to bring attention to the loss of historical information and evidence, due to the disappearance and replacement of websites and documents on the web, and the turnover or deletion of bulletin board system (BBS) newsgroups on the internet (Gardner, 1997, p. 3). Moreover, from at least 1994, libraries, archives and cultural heritage organisations have also had concerns about the ephemerality of web content.[4]

2.3.2 Web Archiving

In recent years, there has been much recognition for the historical, cultural, informational, intellectual, social, political, journalistic, commercial, and evidential importance of archiving web content (Reyes Ayala, 2013; Brügger, 2018; Milligan, 2019; Dougherty, 2007; Schneider et al., 2009; Weigle, 2018; Foot & Schneider, 2006; Ben-David, 2021; Cowls, 2013; Winters, 2017; Weber & Napoli, 2018; Xie et al., 2013; Denev, et al., 2009; Eltgroth, 2009; Taylor, 2017b). This was not always the case (Masanès, 2006, p. 2). Early on, various schools of thought debated to what extent the web should be archived. Arguments against an excessive approach to web archiving relate to editorial quality issues, in so far as the quality of content on the web was inferior to that which had been appraised through a traditional editing panel (Masanès, 2006, p. 6). For instance, Chakrabarti et al. (1999) noted that a web page might contain "truth, falsehood, wisdom, propaganda or sheer nonsense" (p. 54). Capturing everything also presents issues such as: who has the economic responsibility to collect and preserve the web, to invest in the technology to do so, to provide the storage and preservational maintenance, and to provide the finance for research, development, and training? (Lyman, 2002, p. 39; Grotke, 2011; Taylor, 2011). Because of the enormity of the task, it is at least unreasonable, and probably impossible to expect any one institution to assume full responsibility for archiving everything on the web. Thus, a multi-agency worldwide approach has materialised (mostly in developed countries) (Gomes et al., 2011) whereby different institutions in different countries endeavour to capture and preserve

[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, 2017b, 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., 2009, p. 206).

what they can, and what they deem as relevant for their collection mandates and stakeholders. Such institutions include national and regional libraries and archives, university libraries and academic institutions, non-profit organisations, and commercial organisations (Wikipedia, 2011+, List of web archiving initiatives).

Another debate arose within the community of computer scientists who portrayed the web as "a self-preserving medium" (Masanès, 2006, p. 6). Although several studies emerged which would challenge this notion, as the rationale for archiving the web was further reinforced by studies that examine link rot, web content drift, and the extent and frequency of web content change over time. Such studies have been conducted for almost three decades, and span across different disciplines such as education, law, library and information science, information science and technology, computer science, and medical sciences (Harter & Kim, 1996; Koehler, 1999; Lawrence & Giles, 1999; Germain, 2000; Cho & Garcia-Molina, 2000; Lawrence et al., 2001; Markwell & Brooks, 2002; Dellavalle et al., 2003; Fetterly et al., 2003; Hester et al., 2004; Ntoulas et al., 2004; Sellitto, 2005; Goh & Ng, 2007; Wren, 2008; Klein et al., 2014; Zittrain et al., 2014; Zhou et al., 2015; Jackson, 2015a; Bansal & Parmar, 2020; Craigle et al. 2022).

Link rot, also known as reference rot, broken links, or link decay, is used as a term to indicate that a URL no longer provides direct access to a file or web page as originally indicated. Furthermore, even if a URL is stable, the contents of a web page could change; hence, ensuing readers may not view the exact same cited content, or even have the same user experience (Lawrence et al. 2001, p. 30; Dellavalle et al., 2003, p. 788; Schneider et al., 2009, p. 205; Brügger, 2010, p. 6). For example, computer scientists from Stanford University monitored 720,000 web pages daily over a four-month period and found that 40% of web pages in the .com domain changed their web content daily, while web pages in other domains were at an average of 10% (Cho & Garcia-Molina, 2000, p. 201). Another study by Andy Jackson, examined the URLs of web pages that were archived by the UK Web Archive in 2013 and 2014 to see whether such pages were still available on the live web, or had changed. Jackson suggests that "very few archived resources are still available, unchanged, on the current web. After just two years, 60% have gone or have changed into something unrecognizable" (Jackson, 2015a).

In terms of reference rot, a study by Spinellis (2003) used two computer science journals to source a sampling of publications from 1995-1999 which cited URL references and extracted 4,375 URL references for verification. Spinellis found 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). Craigle et al. (2022) describe how link rot “can be a particularly frequent occurrence for law review articles because the law review societies that publish them have not yet adopted standards for preserving online access to them, particularly the adoption of a standard for implementing persistent URLs” (p. 93). Bansal and Palmer (2020) examined “the accessibility, deterioration and half-life of URLs of web documents cited in Current Science Journal published during 2015-2016.” Out of a total of 1724 URLs cited in 1564 articles they found that “little more than half of the citations (56.67%) were active”, while the “rest were found inactive or not working (43.33%)” (Bansal & Palmer, 2020).

For some commentators, the instability of URLs over time may be due to software and system upgrading, changes in filing systems and file names, the re-arrangement of web content, and relocation of servers or server name changes (Berners-Lee, 1998; Besser, 2000; Lyman, 2002, p. 38; Lawrence et al., 2001, p. 28; Spinellis, 2003, pp. 72–73; Pennock, 2013, p. 3; Masanès, 2006, p. 7). Other reasons may be due to the relocation of researchers due to an institutional change (Bansal & Palmer, 2020, p. 1), and the fact that organisations often lose interest in maintaining sites or have not got the time or financial resources to keep sites and URLs up to date (Weisbard, 2011, p. 14). Tim Berners-Lee, the inventor of the web, suggests that link rot occurs often, and more simply due to a lack of human “forethought” (Berners-Lee, 1998).

In 1997, Michael Cunningham of *The Irish Times* described the early web as being somewhat

> likened to one vast, rapidly fluctuating library (of bits rather than atoms). But unlike a traditional library it is being rebuilt every minute. Its sites can flicker and die in days, hours or even seconds. Web pages are revised and spiced up with fancier graphics and revamped designs, more "plug-in" animations and "applets", often with no record kept of the previous mutations (Cunningham, 1997b, p. 18).

In 2003, David Worlock of Electronic Publishing Services Ltd. in London claimed that 25% of the 2,483 British government websites change their URL every year (cited in Weiss, 2003). At the time, Worlock contended that it was problematic as some government documents only existed as a web page, an example being that the dossier produced by the British government on Iraqi weapons only ever appeared as a web page. Thus, Worlock suggests that there is “no definitive reference where future historians might find it” (cited in Weiss, 2003). More recently, Brügger (2018) remarks that at the time of the inauguration of Trump,

there were substantial changes to the official White House website (www.whitehouse.gov), including the removal of topics on climate change and global warming which had been published on the site by the former President, Barack Obama (p. 1). Paul Koerbin of the National Library of Australia also refers to the transient nature of websites during election campaigns, changes of government, party leadership challenges and government leadership changes (Koerbin, 2013a; Koerbin, 2013b). In the Republic of Ireland each change of government leads to changes in departmental titles and often major reallocation of ministerial responsibilities, and up until recently, changes in departmental titles usually entail the creation of new URLs for departmental websites (Healy, 2016; see Table 5.1 in chapter 5.0).

Timestamp: 1996-12-24 (www.irlgov.ie)

Timestamp: 1997-01-05 (www.irlgov.ie)

Timestamp: 2000-03-02 (www.irlgov.ie)

Timestamp: 2002-03-28 (www.irlgov.ie)

Timestamp: 2008-12-07 (www.gov.ie)

Timestamp: 2011-07-03 (www.gov.ie)

Figure 2.2: The changing nature of the Government of Ireland website captured in the Wayback Machine from 1996 to 2008 (www.irlgov.ie), and 2008 to 2011 (www.gov.ie)

In December 2017, the Irish government decided to migrate all the departmental websites under one main website. Starting with the website of the Department of An Taoiseach, departmental websites began migrating to the new centralised website in 2019. The Taoiseach at the time, Leo Varadkar (Fine Gael) outlined the purpose of the plan in Dáil Éireann, as follows.

> Departments are currently represented online by multiple distinct websites and platforms, each providing different visual styles and user experience. A Government decision was taken in December 2017 to migrate all primary Department websites to one single portal, gov.ie. This aligns with international best practice. Gov.ie has been developed with the citizen at its centre, with an emphasis on policy and service areas, as opposed to how a Department is structured internally.

Varadkar further outlined how the users of the taoiseach.ie website were given notice in November 2018, that the information on the site would be migrated to the proposed gov.ie website, and remained available until February 2019. It was also "archived in co-operation with the National Library of Ireland" (Leo Varadkar, Dáil Éireann, Departmental Websites, 26 February 2019). The decision to centralise government department websites into one main website caused some debate on why the government had not done more market research into the decision and surveyed the users to see what they wanted (Micheál Martin, Dáil Éireann, Departmental Websites, 26 February 2019). Nonetheless, it is reassuring to see that the government liaised with the National Library of Ireland to archive the websites before they started the migration. One would also hope that the government will also liaise with the library regarding the "archivability" of their new website. Stanford Libraries describe archivability as "the ease with which the content, structure, functionality, and front-end presentation(s) of a website can be preserved and later re-presented, using contemporary web archiving tools" (Stanford Libraries, n.d., Archivability).

Research by Gomes et al. (2011) provides an overview of global development in web archiving initiatives. This work also provides the base for a Wikipedia article. The article is regularly updated to document the growing number of global web archiving initiatives. It shows a significant increase in web archiving activities by non-profit organisations, commercial organisations, academic institutions and national and regional heritage organisations and associations, with much of the growth occurring in North America and Europe. Some of the earliest efforts in establishing web archiving initiatives may be attributable to the National Library of Canada from 1995, the Internet Archive from 1996, the National Library of Australia (PANDORA) from 1996, the Smithsonian Museum in 1996, and the Royal Library of Sweden from 1997 (Webster, 2017b, pp. 176–178; Brown, 2006, pp. 9–11; Koerbin, 2021, p. 24; Milligan, 2019, p. 76; Arvidson et al., 2000). Several national libraries followed suit to develop web archiving programmes such as the National Library of New Zealand (1999), the National Library of the Czech Republic (2000), the Library of Congress (2000), the National Library of Korea (2001), the National and University Library of

Croatia (2004), and the National and University Library of Iceland (2004) (Wikipedia, 2011+, List of Web Archiving initiatives).

It is also worth mentioning the work of the International Internet Preservation Consortium (IIPC). The IIPC was founded in 2003 by twelve members, being the national libraries of Australia, Canada, Denmark, Finland, France, Iceland, Italy, Norway, Sweden, The British Library (UK), The Library of Congress (USA) and the Internet Archive (USA). The IIPC set out to achieve several goals as outlined below, which are extracted 'verbatim' from an archived copy of the IIPC about/index page from 2004 (Figure 2.3):

- To enable the collection of a rich body of Internet content from around the world to be preserved in a way that it can be archived, secured and accessed over time.
- To foster the development and use of common tools, techniques and standards that enable the creation of international archives.
- To encourage and support national libraries everywhere to address Internet archiving and preservation.

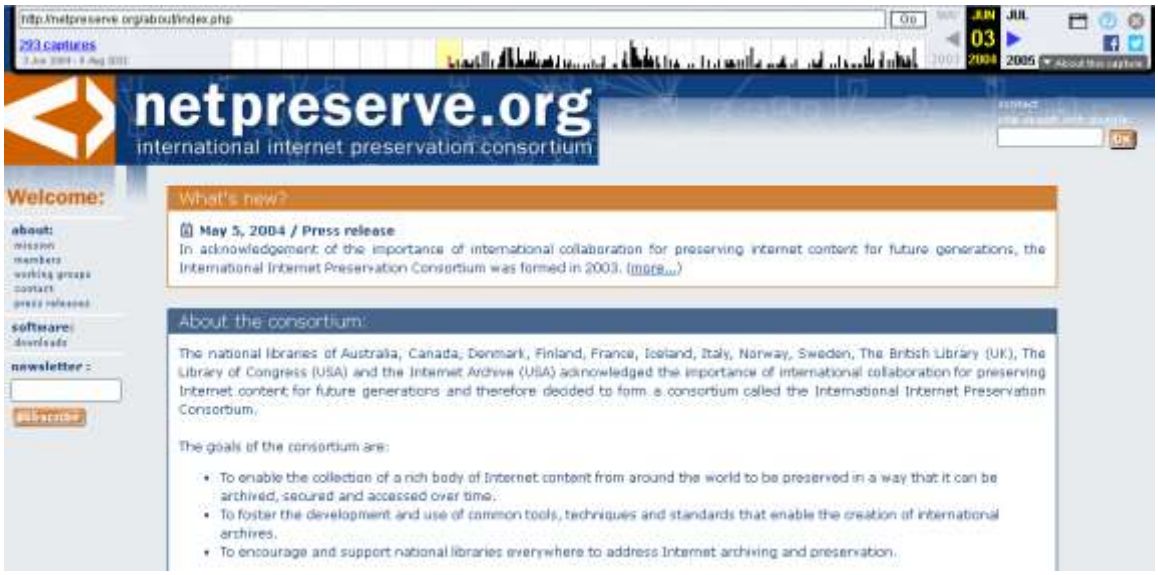

Figure 2.3: Screenshot of the IIPC about/index page, captured in 2004 in the Wayback Machine (Timestamp: 2004-06-03 01:41:15) [5]

Today, the IIPC has over fifty members from thirty-five countries, and has broadened its goals over the years to further include:

- developing "international advocacy for initiatives and legislation",

[5] IIPC about/index page, Wayback Machine, 2004, https://web.archive.org/web/20040603014115/http://netpreserve.org/about/index.php

- fostering “broad international coverage in web archive content through outreach and building curated collaborative collections”, and
- encouraging and facilitating “research use of archived Internet content” (IIPC, n.d., About Us).

The IIPC curated collaborative collections are hosted on the Archive-It platform, and cover topics such as the Summer and Winter Olympics, the Summer and Winter Paralympics, the European Refugee Crisis, and COVID-19 (IIPC, Archive-It, https://archive-it.org/home/IIPC). Also, the IIPC and their members responded to the events taking place in the Ukraine, and through a collaborative effort they have developed a collection around the Ukrainian war (IIPC, n.d., IIPC webinar: Web Archiving the War in Ukraine). This is a good example of how political conditions might influence the selective processes for a web archive collection. The IIPC organises an annual web archiving conference which is one of central events in the calendar of the web archiving community and has seen a dramatic rise in the contribution by individuals who use the archived web for research and is evident from their programme schedule from May 2022 (IIPC, WAC 2022 programme).

2.3.3 Loss of Digital Heritage

While web archiving offers a solution to capture and preserve national digital heritage on the web, it is often the case that governments and societies have not quite grasped the reality of the consequences for the loss of this heritage and therefore lack the urgency to develop strategies for its collection and preservation. The focus is on the creation of new information and not the preservation of the old (Lyman, 2002, p. 39). Indeed, Morris (2019) suggests that the loss of digital heritage is often due to realising “too late that the latest technology is neither mundane nor permanent” (Morris, 2019, p. 500). But, even if the concerns for the preservation of digital heritage on the web is acknowledged, it might take “decades until the technical, organisational, and economic conditions are in place to launch preservation initiatives” (Brügger, 2019, p. 16). We might also add political and legal “conditions” to this list, as they present further barriers for the launch of preservation initiatives and will be discussed in more detail in chapter 4.0 and chapter 5.0. However, to note here, the challenges with the long-term preservation of digital heritage are not unique to the internet or the web.

For UNESCO (2003) some of the factors which contribute to the loss of digital heritage to posterity include technological obsolescence of hardware and software, availability of resources, and “the lack of supportive legislation” (UNESCO, 2003). There is also the case

that “Attitudinal change has fallen behind technological change” and consequently, the “threat to the economic, social, intellectual and cultural potential of the heritage - the building blocks of the future - has not been fully grasped” (UNESCO, 2003).[6] Lyman (2002) notes how societies have lost important parts of their cultural heritage in the past because it was not archived or preserved due to “cultural”, “technical”, “economic”, and “legal” problems (pp. 38–39). Lyman (2002) points out that this is

> in part because past generations did not, or could not, recognize their historic value. This is a cultural problem. In addition, past generations did not address the technical problem of preserving storage media—nitrate film, videotape, vinyl recordings—or the equipment to play them. They did not solve the economic problem of finding a business model to support new media archives, for in times of innovation the focus is on building new markets and better technologies. Finally, they did not solve the legal problem of creating laws and agreements to protect copyrighted material yet at the same time allow for its archival preservation. Each of these problems faces us again today in the case of the Web (pp. 38–39).

While Lyman may have published this summary in 2002, it is still applicable today, although one could argue that the web archiving community has come a long way in providing solutions to the “technical” problem. On the other hand, web archiving is complicated by “ever-evolving” internet, web, and software technologies, thus, such technologies “will always be ahead of the capture tools” (Truman, 2016, p. 20). Nonetheless, continual efforts are being made by heritage organisations and web archive curators to capture what they can, as best they can (Laursen & Møldrup-Dalum, 2017, p. 220).

2.3.4 Legislative Changes & Web Archiving

As a result of emerging publishing technologies several countries began to review and amend their copyright and legal deposit legislations 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. For example, in Denmark, legal deposit for print publications has existed since 1697, through a Royal Ordinance, which mandated publishers to deposit five copies of everything they printed in the Royal Library

[6] With this in mind, the National Library of the Netherlands are responsible for collecting and curating websites hosted by the former Dutch internet provider XS4ALL, to become the “the first born digital collection in the world” to receive a place on UNESCO’s register for Documentary Memory of the World (Teszelszky, 2022).

(Dupont, 1999, p. 244). Thereafter, the law was updated on several occasions for the inclusion of maps, and an extension of legal deposit to include Danish territories, for example. A new law passed in 1997 brought with it "a revolution in the history of legal deposit" as it requested "not only printed material, but all works published in Denmark, regardless of the medium used for the production of copies" including "published works on the Internet, that form a final and independent unit and which have been produced for a Danish audience" (Dupont, 1999, p. 245). The legislation was further revised in 2004 to broaden the scope for the inclusion of archiving the Danish national web domain (Webster, 2017b, pp. 179–180). Web archiving is therefore a legal obligation on the part of some, but not all legal deposit institutions. For example, in the Republic of 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 by the National Library of Ireland and other nominated legal deposit libraries. However, the legislation does not allow for the routine archiving of the Irish national web domain (Ryan et al., 2022). The collection of the web space of Northern Ireland is covered by UK legal deposit through The Legal Deposit Libraries (Non-Print Works) Regulations 2013 which enables the routine archiving of the UK national web domain. This will be discussed in more detail in the next chapter.

While the histories of copyright, legal deposit, and even censorship are often intertwined, they have had different trajectories and interpretations. For example, the concept of copyright only evolved following the invention of the printing press and the spread of its use, as publishers-printers were the first to feel the need to protect their rights to publish a specific work by an author, and prohibit other printers replicating the work, to sell at a lower price (Matthews, 1890, pp. 587–588). The 'concept' of a legal deposit scheme for print publications was first implemented in France in 1537 through the *Ordonnance de Montpellier*, a royal decree by King François I of France. The decree prohibited the sale of any book prior to a copy being deposited in his castle (Larivière, 2000, p. 6; Vène, 2015). While François I was known for his appreciation of books and the arts (Partridge, 1938, p. 5), Vène (2015) suggests that the aim of the decree was to identify works for the preservation of memory, but also to control the dissemination of dissident ideologies (particularly religious ideas). Partridge (1938) notes that the idea of collecting a rich body of literature, at no cost to the state soon spread to other monarchies (p. 3).

In Great Britain, the first implementation of the concept was introduced in 1610 by Sir Thomas Bodley, the founder of the Bodleian Library at Oxford University in 1602. Bodley negotiated a deal in 1610 with the Company of Stationers (London), for them to supply a copy of every new book to the Bodleian Library, published by its members (Muir, 2005, p. 1;

Larivière, 2000, p. 6; Feather, 1994, pp. 97–98). Similar efforts in the concept of deposit schemes for print publications were introduced in Sweden (1661), Poland (1645), Denmark (1697) and Finland (1702) (Larivière, 2000, p. 7; Muir, 2005, p. 9). The development of more elaborate legislations regarding the copyright of print publications, and by extension legal deposit, evolved over the centuries on a national basis.

In general, most legal deposit schemes require producers of print publications to deposit a copy of each new publication in a nominated institution, often designated as a national library, state library or university library (Gooding et al., 2019, p. 9). It serves as a system to compile, preserve, and provide access to a comprehensive collection of a country's publication outputs, providing a significant contribution to national cultural heritage. For Lariviere (2000),

> Legal deposit legislation serves a clear national public policy interest by ensuring the acquisition, the recording, the preservation and the availability of a nation's published heritage. Such a national collection is undoubtedly one of the major components of a country's cultural policy and should also be considered as the foundation of a national policy of freedom of expression and access to information (p. 4).

While legislation in some countries has allowed for 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).

2.4 Summary

This chapter focused on the main causes for the loss of digital heritage, including Irish digital heritage (RQ1). It further examined some of the challenges for participation in web archive research (RQ2), and the literature offered some insights on how to improve the conditions for conducting web archive research (RQ5). First, the chapter positioned heritage within the broader framework of societies, communications, and culture, and argued that it is within the intersections of these concepts that heritage is produced. In doing so, it provided an understanding of a society as a large social group, made of individuals, who interact and communicate, who have multiple things in common (e.g., territory, language, traditions, culture, political institutions), and have some levels of consciousness that they differ from other societies. It offered some insights on what constitutes the heritage of a society, and how national heritage should be inclusive of a society's sub-groups, ethnic groups, and communities. It also highlighted how the advent of the internet, web, and social media generated spaces for alternative discourses that foster a greater pluralism of perspectives, and by-pass traditional modes of media gatekeeping, and has driven the growth of online communities with their own social and cultural values and therefore, also their own heritage. Thereafter, the chapter explored some of the underlying reasons for web archiving, and how this stemmed from wider concerns on the loss of digital heritage in general, with the web just being another media carrier to worry about.

Some of the main causes for the preservation of digital heritage in general were identified such as technological obsolescence, and media deterioration, and simply because attitudinal change has fallen behind technological advancements (UNESCO, 2003). Lyman (2002) also notes how societies have lost important parts of their heritage due to the inability of past generations to recognise its importance and historic value, and due a lack of foresight and technical ingenuity to ensure continuity for preservation, storage, and maintenance (pp. 38–39). Furthermore, Lyman (2002) posits how the economic problem stems from the failure to find a business model to support the archiving of new media formats, while the legal problem stems from the failure to create legislation which protects copyright while at the same time allows for archival preservation (p. 39). These problems equally apply to the loss of digital heritage on the web, and Ireland is no exception.

The chapter documented how there have been continual warnings to the ROI government regarding the loss of digital heritage and electronic records due to obsolete technology, and the fact that there is a lack of formal record keeping guidelines for government departments and agencies for electronic records. The literature illustrated how the evolving nature of

publishing in the past 50 years became problematic for legal deposit legislation which was fundamentally print-centric. As a result, several countries began to amend their copyright and legal deposit legislation from the 1990s to incorporate the deposit of non-print materials, as well as the incorporation of the web archiving of a country's national web domain, as a matter of routine. Although as mentioned, while legal deposit legislation was updated in the ROI in 2019, it failed to include the routine web archiving of the Irish national domain as part of the national legal deposit scheme, and so the ROI continues to have mass losses of Irish digital heritage. The chapter demonstrated how web archiving is a necessary activity for the preservation of national digital heritage.

3.0 OVERVIEW OF WEB ARCHIVE RESEARCH

> *In the past, important parts of our cultural heritage have been lost because they were not archived—in part because past generations did not, or could not, recognize their historic value. This is a cultural problem. In addition, past generations did not address the technical problem of preserving storage media—nitrate film, videotape, vinyl recordings—or the equipment to play them. They did not solve the economic problem of finding a business model to support new media archives, for in times of innovation the focus is on building new markets and better technologies. Finally, they did not solve the legal problem of creating laws and agreements to protect copyrighted material yet at the same time allow for its archival preservation. Each of these problems faces us again today in the case of the Web (Lyman, 2002, pp. 39–40).*

The previous chapter positioned digital heritage within the broader frameworks of societies, communications, and culture, and argued that it is within the intersections of these concepts that heritage is produced. It further examined the concerns and main causes for the loss of digital heritage, and how this relates to Ireland (RQ1). This chapter examines how web archiving has emerged as a solution for some of those challenges, however, that is not to say that is an easy solution, rather there are multiple challenges related to these activities.

Using desk research and a literature review from across multiple disciplines, this chapter explores the international literature which describes the challenges for web archive research (RQ2). As outlined in chapter 1.0, this thesis considers web archive research to be representative of the processes and activities described in the Archive-It web archiving lifecycle model which includes appraisal, selection, capture, storage, quality assurance, preservation and maintenance, replay/playback, access, use and reuse (Bragg & Hanna, 2013). Thus, for the purpose of this chapter, web archive research is inclusive of web archiving, curation, and the use of web archives and archived web content for research or other purposes. First, the chapter provides an overview of web archive research, starting with web archiving and curation, and examines some of the challenges experienced by the web archiving community. The next section examines scholarly engagement with web archives and the challenges experienced by this user community. Then, it offers a brief overview of the literature, which is relevant for studying the archived web, and discusses the value of web archives for research or other purposes. The final section provides a review

of the literature on web archive creators and users, combining studies on web archiving practices, web archive users and scholarly engagement.

3.1 Web Archiving and Curation

It is widely agreed that web archiving involves the 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 have 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 compared to archiving a static or semi-static web page, as well as different legal, ethical, and curatorial considerations (Breed, 2019; Bingham et al., 2020; Bingham & Byrne, 2021; 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 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, 2020a).

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. For Aasman et al. (2019) a "web archaeological" approach focuses on

> actively uncovering the history of the web in its early days, emphasising the role of 'digging' and 'reconstructing' as central methods in tracing material objects (software, hardware, terminals, hard drives, cables, et cetera) and born-digital objects (websites, web elements like banners or avatars, blogs and vlogs, and many other forms of user-generated content) (p. 2).

Also, 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. 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 (Alberts et al., 2017; de Haan et al., 2017).

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, 2022a; Jackson, 2022b; 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. For Truman (2016) this is inclusive of "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, one should also consider the development and implementation

of software and tools as a necessary part of the activities and processes undertaken in web archive research.

Search and retrieval capabilities have also presented challenges, and while the web archiving community has worked on improving its search capabilities through complementing traditional URL search with metadata and full-text search, they have encountered considerable challenges along the way. URL search entails the input of a URL, which means the URL needs to be known in the first place. Metadata search entails a search through "metadata attributes, such as category, language and file format" (Costa, 2021, p. 72–73). However, the "manual creation" of descriptive metadata for selective curated collections is "a time-consuming and expensive process". This makes it "a non-viable option" for large-scale web archives (Costa, 2021, p. 72). Costa (2021) notes that metadata needs to be created automatically for large-scale collections, 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 (Costa, 2021; Holzmann & Nejdl, 2021; Winters & Prescott, 2019; Nielsen, 2016; Jackson et al., 2016). Since web archive search also includes a temporal dimension, algorithms that were developed for ranking search results from the live web will not provide satisfactory results. Jackson et al. (2016b) highlight the difficulties of designing a ranking model that satisfies scholarly requirements, as "some scholars [...] questioned the very idea of relevance ranking" (p. 105). For Jackson et al. (2016) if a ranking model is used it must be made completely transparent so scholars can interpret the results accordingly.

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 Dictionary of Archives Terminology 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" (Society of American Archivists, 2005+, Provenance). 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, 2022b). 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 (2017), 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, 2021c). 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).

Moreover, the literature illustrates how 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 (Winters, 2020a; Hock-Yu, 2014; Gooding et al., 2021; Vlassenroot et al, 2019; Graham, 2019; Ogden, 2021; Brügger, 2021c; Ogden et al. 2022; Ben-David, 2021).

As mentioned in Chapter 2.0, many countries have legal deposit legislation which enables the archiving of non-print publications such as e-books, audio books, and e-journals, as well as the archiving of the country's national web domain. Archiving a national domain tends to include a crawl of a country code top level domain (ccTLD), without having to request permissions from website owners. However, defining the scope of a national domain is problematic. For example, if a domain web archive is based solely on a ccTLD like .ie or .uk, it will inevitably exclude thousands of websites with domains such as .com, .org, or .net (Day, 2003, p. 16). Day (2003) further suggests that national domain archives might also include websites from servers that are physically located in a country, or websites that belong to organisations or individuals when "the intellectual content is of relevance" for national digital heritage (p. 16).

Access to national domain legal deposit collections is also dependent on the legislation, and for the most part, may only be accessible onsite in a designated building or reading room.

There are some exceptions such as the Croatian and Icelandic national domain web archives, which are both available online as open access (Winters, 2020a, p. 160). On the other hand, while the Swedish legislation allows for the collection of the Swedish domain, there is currently no provision for access, and the Danish web archive can only be accessed offsite by researchers with support from an academic institution on a project permissions basis (IIPC, Legal Deposit, n.d.; Winters, 2020a, pp. 160–163). Web archiving initiatives that engage in legal deposit collection tend to combine their collection efforts with selective collection which seeks permissions from website owners for capture, as well as to provide access to the archived website in a public web archive. Countries that do not have the necessary legal deposit legislation for archiving national domains, have few options but to engage in collection development based on fair use, or through permissions-based selective collection only, which is a resource intensive process (Costa, 2021, p. 72).

There are other challenges with permissions-based selective collections. Not all websites provide contact details and even if a contact is found there is no guarantee that a website owner will respond (Ryan et al., 2022; Bingham & Byrne, 2021). Byrne (2020b) highlights the challenges of identifying contact details to request open access permission for content archived by the UK Web Archive, whereby, even when contact details are identified and permission requests are sent to content owners, there is a very high failure rate. On average only 20% of requests sent by the UK Web Archive have resulted in open access permission being granted. However, this is a general figure across the whole archive. The actual figure for individual curated collections varies depending on the type of content that is selected. As of June 4, 2018, the response rate for the *Sport: Football* collection was at 10.49% (Byrne, 2020b, p. 5).

Pennock (2013) further discusses a weakness of selective web archiving due to "the possible or unintentional and unacknowledged selector bias" (p. 9). Pennock (2013) explains that the selection of websites is

> commonly a manual process that reflects the particular interests or knowledge of the person(s) choosing sites for the collection. The sheer size of the Internet, the number of websites hosted and the speed at which information can be published, all make it very difficult for manual selectors to keep abreast of new sources, especially for event-based collections (p. 9).

Brown (2006) further points out that "the greater the degree of selectivity employed, the more subjective the resultant collection will be, constraining the as-yet-unknown requirements of future researchers" (p. 32). As already noted in Chapter 2, the very fact that

the concepts of in-groups and out-groups are acknowledged as a phenomenon of human behaviour will also influence selector bias for what is included or excluded as part of a selective thematic collection. For example, selector bias may affect the inclusion of representations from ethnic minorities, societal sub-groups or communities who go against the grain of preferred meanings of a dominant hegemonic social group. This is why legal deposit libraries tend to conduct both selective and domain-wide web archiving as a more balanced, representative, and inclusive approach towards the capture of national digital heritage on the web.

Other global web archiving initiatives archive the web on a fair use basis, regardless of borders, such as the Internet Archive, a non-profit heritage institution based in San Francisco. Rated as one of the largest web archives in the world, the Internet Archive began web archiving in 1996, and currently provides "unrestricted access" to their web archive through the Wayback Machine (Webster, 2017b, pp. 176–177; Brown, 2006, p. 9). The Wayback Machine, as a search interface, was not publicly available until 2001 (Rogers, 2013, p. 65). The Wayback Machine is globally accessible online and allows users to save a web page by inputting a URL to be captured in real-time. While the Internet Archive collects on the grounds of fair use, it provides an opt-out takedown clause for website owners who do not want their websites available in the public Wayback Machine (Lowcock, 2020). The clause states that: "The Internet Archive may, in appropriate circumstances and at its discretion, remove certain content or disable access to content that appears to infringe the copyright or other intellectual property rights of others." Website owners then need to provide relevant details to the Internet Archive inclusive of a statement "made under penalty of perjury" that the information they provide "is accurate" and that they "are the owner of the copyright interest involved or are authorized to act on behalf of that owner" (Internet Archive Help Center, n.d., Wayback Machine General Info).

3.2 Web Archives and Scholarly Engagement

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 alone (Gomes et al., 2021b; Brügger & Laursen, 2019; Brügger & Milligan, 2019; Brügger, 2017; Brügger & Schroeder, 2017). Nonetheless, several commentators observe how scholars were slow to engage with web archives as a research resource (Webster, 2020; Rogers, 2019; Leetaru, 2019; Webster, 2017b; Winters, 2017; Leetaru, 2017; Meyer et al., 2011; Dougherty et al., 2010). The next

section examines some of the reasons for the slow development of scholarly engagement with web archives.

3.2.1 Challenges for Scholarly Engagement

There are several reasons put forward for the lack of scholarly 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; Winters, 2017; Costea, 2018). Initially, web archiving initiatives tended not to prioritise how web archives would or might be used (Dougherty et al., 2010, p. 10; Hockx-Yu, 2014 p. 113; Schroeder & Brügger, 2017, p. 12; Gooding et al., 2021, p. 1163). Rather, Thompson (2008) suggests web archiving institutions "followed a traditional model of acquisition where material is held in the belief that it has value even though there may be no immediately identified user" (pp. 19–20). For example, the New Zealand National Library did not examine the extent of awareness and use of their web archive before late 2014, even though they commenced a web archiving initiative in 1999 (Riley & Crookston, 2015, p. 3). Other surveys of the procedures, practices and policies of web archiving institutions show similar tendencies. The National Digital Stewardship Alliance, Content Working Group (2012), found "an area of uncertainty" by web archiving institutions vis-à-vis how collections were being used (p. 11). Schroeder and Brügger (2017) also note that for many years, web archiving initiatives struggled

> to set up archiving procedures, hardware and software to keep pace with the seemingly endless flow of new web content and ever evolving software development, while little attention was paid to who might use the material in the archive, and how it might be used (p. 12)

In a Harvard Library report, Truman (2016) stresses the need "for greater communication and collaboration" with researcher communities, as well as the "the need to gather researcher feedback on requirements and impediments to the use of web archives" (p. 42). Although, to do this, one might need to identify a community of users, or even potential users in order to attain feedback (c.f. Ras & van Bussel, 2007; Stirling et al., 2012; Gooding et al., 2019). Thus, Truman (2016) identifies the need for more communication and collaboration between those who curate, create and steward web archives and those who use (or might use) a web archive for purposeful research (p. 3). Certainly, it could be argued that a lack of dialogue or collaboration between the creators of web archives, and end users (or even potential end users) has had some effect on engagement with web archives for research purposes. This needs to be addressed going forward.

On a positive note, collaboration between web archive creators and end user researchers has been improving over the past decade (Schroeder & Brügger, 2017, p. 12–13; Webster 2017b, p. 187; Maemura, 2022, p. 6). This is partly due to growing efforts to foster and increase research engagement by consortiums, networks, research projects and libraries in some instances. Examples include: the International Internet Preservation Consortium (IIPC), the Research Infrastructure for the Study of Archived Web Materials (RESAW), the Analytical Access to the Domain Dark Archive (AADDA) project, 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. In addition, the Archives Unleashed project launched a programme to provide support for cohorts, to further foster research engagement (The Archives Unleashed Project, n.d., Archives Unleashed Cohorts), while Arquivo.pt developed an annual award for initiatives which use and demonstrate the use of the Arquivo.pt web archive (Arquivo.pt, n.d., Arquivo.pt Awards).

The conference programmes of organisations like the IIPC and RESAW often feature workshops for end users and researchers who engage with web archives. The UK Web Archive also provides support for researchers and PhD students using its collections and has been proactive in collaborating on workshops and training. Indeed, the IIPC, libraries and other like-minded organisations often collaborate to provide seminars and workshops for web archive users/researchers, some of which include working with big digital data. When it comes to handling such data, there is often a prerequisite to have some programming skills. For example, in a call for participation of researchers at an Archives Unleashed 4.0: Web Archive Datathon held in the British Library in June 2017, it was suggested that: “Researchers should be comfortable with command line interactions and knowledge of a scripting language (such as, but not limited to Python) is strongly desired” (IIPC members list, Email, 04 April 2017). While this is certainly a good thing for those who are comfortable with programming, it might also be seen as a barrier for entry by scholars with limited technical skills (Bingham & Byrne, 2021, p. 5).

There is also the need to consider that some academics are more comfortable with, and trusting of, ‘proven’ traditional research methods, although this is not something unique to web archive research. Other disciplinary fields that have had some history of engagement in the use of computational methods for big data analysis have had similar stories with traditional researchers being mistrustful of computing methods. Examples include

humanities computing, history and computing, and the development of computational methods in the social sciences (Drucker, 2012; Hindley, 2013; Winters, 2018; Kelle, 1997). Indeed, in the social sciences, it was only with the rise of the personal computer in the 1980s that computational methods for qualitative research began to gain any kind of traction in the academy, despite the availability of software and tools since the late 1960s (Kelle, 1997, pp. 2–3).

Challenges arise, due to the characteristics of an archived website or web page which may not be a complete surrogate of what was once on the live web, rather, it is a version (Brügger, 2010, pp. 6–7). Brügger and Finnemann (2013) propose that the archived web is "a Reborn, Unique and Deficient Version and Not Simply a Copy of What was Once Online" (p. 74). Deficiencies in the archived artefacts may occur because of the temporal dimensions such as the time it takes to capture, and the possibility of content updates during capture. Deficiencies may also occur due to technical issues such as glitches during the archiving process such as robots.txt or limitations with the archiving software/hardware to keep up with the constant change and upgrade of web media file types and the evolving nature of dynamic content (Brügger, 2010, p. 7; Meyer et al., 2011, p. 6; Pennock, 2013, p. 13; Maemura, 2018, p. 332; Bingham & Byrne, 2021, pp. 3–4).[7] For example, Morris (2019) and Aasman (2019) discuss the absences of sound and audiovisual content in web archives. Aasman (2019) notes how the Wayback Machine is "unable to reproduce flash-based videos" and thus, early captures of YouTube pages "show nothing but a front page with empty screen" (p. 43). Morris (2019) also draws attention to the challenges for finding sound files in a web archive. First, while there may be an icon displayed for a sound file on an archived webpage, the actual audio file may not have been captured (p. 497). Second, "preserving audio formats often require preserving the sounds themselves as well as the technologies on which to play those sounds", thus, even if the file is there, there may be no way to open it or play it without the obsolete software which created it (Morris, 2019, pp. 497–499).

In addition to the challenges above, in order to preserve a website or web page in its entire capacity to produce meaning, it should be inclusive of links to external (hyperlink) information, and quite often this is not achieved due to selection criteria, acquisition policies, technical glitches, financial constraints, or legislative and copyright restrictions

[7] Robots.txt, also known as the robots exclusion standard, or robots exclusion protocol, refers to a standard that is used in websites " to indicate to visiting web crawlers and other web robots which portions of the website they are allowed to visit" (Wikipedia, 2002+, Robots exclusion standard).

(Besser, 2000; Milligan, 2019, pp. 45–46; Hockx-Yu, 2014, pp. 114–115). And, if hyperlinks direct to social media sites, this presents an additional set of technical, ethical, and legal challenges (Breed, 2019; Bingham et al., 2020; Bingham & Byrne, 2021; Vlassenroot et al., 2021). Finally, the collected web content may undergo technical processes during collection, preservation and to provide access through replay or playback (Brügger 2016, 2018; Schneider et al., 2009). Thus, for Brügger (2019; 2018; 2016) archived web content may be considered as reborn digital media, which is clearly distinct from other types of archived media such as film, television, photographs, and newspapers. Consequently, Brügger (2018) suggests that "historians have to become familiar with this type of source, its characteristics, and how these characteristics impact its scholarly use "(p. 3). Therefore, this implies that the use of archived web content for scholarly purposes has ongoing pedagogical challenges.

Other commentators describe challenges with web archives due to the differences between searching on the live web, and searching in a web archive (Costa, 2021; Holzmann & Nejdl, 2021; Winters & Prescott, 2019; Jackson et al., 2016; Nielsen, 2016). URL search is offered by most web archives as an entry point to find archived web materials, such as the UK Web Archive, Archive.today, and the Internet Archive's Wayback Machine, requiring that the user knows the URL in the first instance. Alphabetical browsing is offered by a few web archives such as the UK Government Web Archive and the PRONI Web Archive, while other web archives offer 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, however, for large web archive collections, this presents challenges due to the huge amount of query returns, which also have a temporal dimension (Winters & Prescott, 2019, pp. 398–399; Jackson et al., 2016b, p. 105; Nielsen, 2016, pp. 22–23; Costa & Silva, 2010). Furthermore, full-text searching within web archive collections "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" (Healy et al., 2022, p. 8; Winters & Prescott, 2019, p. 398). However, as noted by Healy et al. (2022) search capabilities are also a challenge for web archive creators. It is expected in current web design that search boxes or interactive filters are part of navigating a website on the live web. However, these features in the archived version of the website are often redundant as the web archive crawler can only follow clickable links and cannot replicate the dynamic interactions that are part of the live website.

Legislation on copyright and legal deposit also presents challenges for researchers to utilise web archives. Using the UK Web Archive legal deposit collections as an 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. Such challenges include the locked down nature of the library terminal whereby researchers cannot view the source code, and so, it is useless for studies in the evolution of code/CSS design which is important for "web historiography"; nor can a researcher copy the URL from the browser, which causes problems for citation (Winters, 2020a, p. 164; Milligan, 2015). Users are not allowed to copy and paste text which totally disrupts the affordances that are used by researchers worldwide, when they use the live web as a source for research (Milligan, 2015). Also, users can not take photographs or screenshots of the screen, rather they must pay for a printout of an archived web page, which is ironic, as researchers are allowed to use cameras to take photographs of historical documents in most archival environments (Milligan, 2015). Furthermore, no two people can view the same instance of an archived web page simultaneously, even if viewing the same content at different Legal Deposit Libraries, which inhibits collaborative research as well as the use of the resource for teaching in the context of classroom group projects (Winters, 2020a, p. 164, Talboom, 2022). Such challenges are manifested due to the restrictive nature of the UK legal deposit legislation as laid out in The Legal Deposit Libraries (Non-Print Works) Regulations 2013 (NPLD).

Gooding et al. (2019) offer other examples for the challenges with the NPLD access protocols. For instance, they discuss how the disciplines like digital humanities, data sciences, and quantitative social sciences have evolved to require "libraries to develop new forms of licencing, collection management and support for digital materials in response to user needs" and how the UK government has supported "computational research through a 2014 copyright exception that allows non-commercial text and data mining of copyrighted materials" (Gooding et al., 2019, p. 7). Nevertheless, they highlight how this sentiment is not extended to NPLD collections, as the NPLD regulations "make no allowance for text and data mining, or to allow materials to be made accessible at the end of their copyright term" (Gooding et al., 2019, p. 7). They suggest that the lack of planning for text and data mining is "now a significant barrier for innovative research" (Gooding et al., 2019, p. 24).

Another problem relates to how government reports emphasise "inclusion and access" and how "scholarly publishing is increasingly transitioning towards Open Access" which is also supported by government and research initiatives, and they highlight how "copyright regulations have been enhanced to allow the provision of accessible copies of materials for readers with a recognised disability" (Gooding et al., 2019, p. 7). Yet, aspects such as these

> are not formally reflected in the NPLD regulations, which use as a basis the Copyright, Designs and Patents Act 1988 (1988) as amended by the Copyright (Visually Impaired Persons) Act 2002 (2002). This means that the 2013

> regulations only allow for accessible copies of NPLD materials to be made available for readers with visual disabilities, rather than all persons with a recognised disability. As such, there is a gap in understanding of the extent to which NPLD supports emerging practices relating to Open Access and accessibility for disabled readers (Gooding et al, 2019, p. 19).

Elsewhere Gooding et al. (2021) suggest that the user was neglected as a stakeholder when it came to drafting the legislation for NPLD access protocols, which is fundamentally print-centric, despite the digitality of the resources being preserved for future generations. Moreover, they insist that because the NPLD ethos is print-centric, it fails to consider the user in line with digital user expectations, and current trends in information seeking behaviours (Gooding et al., 2021). Therefore, while some legal deposit schemes might allow for the collection of websites, they may not effectively deal with the provision of access (Healy et al., 2022). For example, Maurer (2022) notes how the provision of onsite 'only' access to web archive collections in a designated building makes web archives geographically inaccessible for many researchers. Maurer (2022) further suggests that with these types of "closed" archives there is usually very little data

> publicly available about their contents, so it is difficult to convince researchers to travel to the reading room when the researcher doesn't know in advance what exactly the archive contains and whether it's pertinent to their research question.

Therefore, one could argue that these types of access conditions present barriers for innovation.

Truter (2021) also highlights the challenges for end user researchers in terms of the access and use of archived web content 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). This manifests challenges for not only the use of the data, but also affects how and if the data can be made shareable and reusable (Truter, 2021) and runs counter to the requirement of open science which is being stipulated by a growing number of research institutions and funding agencies (Winters, 2020a, pp. 167–168). However, as pointed out previously, legislation is also a challenge for the creators of web archives, for both its collection, and the conditions for access.

Researchers may also be more interested in using big data methods such as topic modelling or network analysis on a web sphere of websites (WARC files) from a specific web archive collection (e.g., Geocities) or to do a longitudinal study across multiple legal deposit annual

web domain collections (see Milligan, 2019; Brügger et al., 2017; Brügger et al., 2019). However, Maurer (2022) points out that "handling the raw WARC files [is] difficult for all parties, so sending extracts of data from institution to research team is often not feasible." Reasons for this are varied and may be "due to a mix of curatorial, technical, legal, economic and organisational constraints" (Brügger, 2021c, p. 217). The Archives Unleashed project based in Canada, in collaboration with the Internet Archive, sets out to provide some solutions to this issue through the development of tools and solutions for handling and analysing large volumes of WARC files that are hosted on the Archive-It platform. However, this would not apply to heritage institutions that collect web content through, for example, in-house crawling. This is why Brügger (2021c) stresses the need for solid research infrastructures between the web archives with the data and the research teams wishing to use the data, and this will help overcome some of the legal, ethical, and technical challenges for both communities. Of course, this will require funding, and a cultural shift placing the creator and user as partners in the full web archiving lifecycle.

Challenges for researchers also arise due to ethical issues. Graham (2017) argues that there has been little attention paid to "the ethics of experiencing and accessing the past web" (p. 103). For example, Graham (2017) 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 (2017) 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). Maemura (2018) 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). Ogden et al. (2022) suggest that researchers need to be vigilant using web archives when researching socially vulnerable communities and highlight the importance of considering "how particular vulnerabilities can be exacerbated over time when linked to individuals—for example, when researching children, social stigmas (e.g., self-harm communities), or identifying past evidence of illegal activity" (p. 17).

While noting the value of web archives as resources for researching online communities and bottom-up histories, Mackinnon (2021) 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 on the free GeoCities hosting 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).

Other scholars illustrate challenges in using web archives due to political and sociotechnical circumstances. Ben-David (2019) discusses how a ccTLD is delegated to countries which have been recognised by the United Nations (UN), and how "the marked boundaries of these portions of the web comply with the geographical borders that divide nation-states", and notes how "this assumption is grounded in the practice of web archiving at national libraries" (p. 90). However, for Ben-David (2019) this assumption becomes problematic when it comes to studying web histories of countries that do not have a ccTLD, such as Kosovo. While Kosovo declared unilateral independence from Serbia in 2008, and was recognised by 133 countries, it is not recognised by the UN due to a veto by Russia, and so, it does not qualify to be allocated a ccTLD (Ben-David, 2019, pp. 91–92). Thus, a Russian veto in the UN for the "recognition of Kosovo has had immense implications on its official presence as a national web", and consequently presents challenges for tracing Kosovar website histories in a web archive (Ben-David, 2019, p. 95). Ogden and Maemura (2021) examine how the sociotechnical, organisational, and resource constraints "under which most web archiving programmes operate" needs to be understood by researchers and suggest that researchers need to become familiar with the "specific limits and constraints, legal governance frameworks, collection mandates, as well as configurations (i.e., of sub-collections) and terminology used for specific collections" (Ogden & Maemura, 2021). Schafer et al. (2016) also discuss "the multiple socio-technical mediations, arrangements, and agencies mobilised throughout the archiving process – be they technical or human", and suggest that understanding a web archive "implies opening several black boxes" in order to understand "the human and technological decisions which lead to its constitution, as well as the creation of this source which is never an exact copy of the original" (pp. 2–3).

While it may seem obvious that historians, media scholars and social scientists will use web archives as resources to document histories of the 1990s and the early millennium, this is not the case (Winters, 2017, p. 174; Ruest et al. 2021, p. 6). Both Brügger (2016) and Schafer (2019) suggest that web archives present challenges due to the "absence" of a traditional style catalogue or registry as an entry point. Costea (2018) identifies a need for improvements to web archives in the areas of discoverability options, data selection, data

management, and access to more comprehensive documentation and metadata. Winters (2017) points out that a major challenge for historians in

> working with web archives is, quite simply, that it is difficult; it requires skills that many historians do not have, and in the short term may be unwilling to learn; it involves acknowledging a degree of ignorance with which otherwise seasoned researchers may be uncomfortable (p. 174).

Truman (2016) suggests that 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. Ruest et al. (2021) also refer to the challenging nature of using big data from web archives due to the "size on the order of petabytes, billions of words, tens of thousands of images, all with murky metadata, provenance, and difficulty to access" (p. 6). Whereas traditional researchers may want to take a more qualitative approach towards using the archived web, they too have challenges due to a lack of research methods and theoretical paradigms for the use of the archived web (Millward, 2015; also see Table 4.15).

Other challenges relate to the fact that some large-scale web archives, such as the Internet Archive's Wayback Machine, may lack depth and are deemed as too broad to meet the needs of specific research which often requires precise datasets (Schneider et al., 2009, p. 215; Dougherty & van den Heuvel, 2009, pp. 1–5). Therefore, researchers often turn to developing their own web archive collections for their needs (see for example, Foot & Schneider, 2006; Engholm, 2000), and so, the research question sets the tone for the identification of what gets captured and why, as well as for how often and for how long. The research question will also drive the approach and methodology, and thus, various methods may be used for the "active process" of archiving web content, such as saving a screenshot or a PDF of a web page, screencasting the functionalities of a web page, or using an archival crawler to collect WARC files (Brügger , 2018; pp. 81–82). For instance, a researcher may collect WARC files to conduct network analysis or topic modelling, or collect images or screenshots for visual discourse analysis, or collect HTML or CSS files to study the evolution of code over time. However, such collections are often narrow in scope and may never be useful for anything other than the study for which they were created (Dougherty & van den Heuvel, 2009, pp. 6–7). Creating a potential for combining researcher-centric web archive collections and derived datasets into collaborative entities may be one answer, but who takes responsibility for the custody, maintenance, and finance of such an entity? Moreover, as different methods are often used for capturing, storage, preservation, and metadata (or

not), how would it become interoperable across systems? On the other hand, some researchers do turn to the Wayback Machine for single site histories (e.g., Hofheinz, 2010; Bødker & Brügger, 2018) or to develop their own curated collections and derived datasets for code analysis (e.g., trackers, third-party cookies etc.) or network analysis (e.g., hyperlinks) (Rogers, 2019, pp. 51–53).

Subscription based web archiving services are also another option for researchers to self-archive and curate web archive collections. For example, the Internet Archive's Archive-It service offers a subscription service, based on the amount of data that is archived annually (Archive-It , 2021). In addition, some web archiving initiatives accept nominations from the general public either as part of a specific campaign or as part of their mainstream service. Content (if in scope) can be nominated to the UK Web Archive for preservation, and the NLI Web Archive offers an option to "Suggest a Website" as part of their efforts. Also, sites can be automatically self-archived in the Internet Archive's Wayback Machine, Archive.today, and in the Portuguese web archive Arquivo.pt (Table 3.1). These are useful services for researchers who want to ensure that the content they are consulting in their research is archived at the same time (or near to that time, in the case of the UK Web Archive) as they are consulting it. In addition, this material can be preserved in more than one place, making it more accessible to other researchers once the content changes or goes offline (Byrne, 2022).

Table 3.1: Useful self-archiving web services for researchers and other users

Provider	**Service**	**URL**
Archive.today	Webpage capture	https://archive.ph/
Arquivo.pt	Save Page Now	https://arquivo.pt/services/savepagenow
NLI Web Archive	Suggest A Website	https://www.nli.ie/collections/our-collections/web-archive
UK Web Archive	Save a UK website	https://www.webarchive.org.uk/nominate
Wayback Machine	Save Page Now	https://archive.org/web/

Despite the shortcomings described above, web archives contain records and documentary evidence of human society and are gradually being recognised and used as resources for the study of the recent past. This is evident in the growing number of edited collections and monographs published on the topic in the last number of years alone (Gomes et al., 2021; Brügger & Laursen, 2019; Brügger & Milligan, 2018; Brügger & Schroeder, 2017; Brügger, 2017; Milligan, 2019; Brügger, 2018). In addition, there has been a growing number of journal publications, conference papers, and conference presentations which discuss the use of web archives as resources for research, or which offer case studies in the use of web archives and archived content. However, it may be difficult for novices to bring such literature together for reading, as it is spread across various fields of the academy covering topics such as media and journalism, social sciences and ethnographies, public health and telemedicine, information science and law, internet studies and web histories, and more (Schafer et al., 2019; Weber & Napoli, 2018; Bødker & Brügger, 2018; Aust, 2014; Ogden, 2021; Ogden et al., 2022; Gorsky, 2015; Adelmann & Franken, 2020; Cocciolo, 2015; Holzmann et al., 2016; Costa & Silva, 2010; Jatowt et al., 2008; Eltgroth, 2009; Taylor, 2017b; Brügger & Finnemann, 2013; Brügger, 2016; Nanni, 2017; Rogers, 2017; Raffal, 2018; Aasmann, 2019; Mackinnon, 2022; Paßmann et al., 2022).

3.2.2 Studying the Archived Web

Becker (1938) succinctly sums up historiography as "the study of the history of historical study" (p. 20). On one hand, Becker (1938) notes that an objective of "historiography is to assess, in terms of modern standards, the value of historical works" (p. 20). On the other hand, Becker (1938) infers that historiography might be treated

> as a phase of intellectual history […] which records what men have at different times known and believed about the past, the use they have made, in the service of their interests and aspirations, of their knowledge and beliefs, and the underlying presuppositions which have made their knowledge seem to them relevant and their beliefs seem to them true (p. 21).

In explaining web historiography, Brügger (2012) puts forward two distinctions in terms of a web historiography aiming at writing the history of the web, or "a 'web-minded' historiography that is a historiography which pays attention to the role of the web in present day society" (p. 103). In explaining further, Brügger (2012) notes it is worth understanding that, since the development of the web, it has become "an integrated part of historiography since more and more documents are made available on the web, and the web is part of the historian's methodological toolbox" (p. 103). Moreover, the web as a platform also informs

"contemporary historiography" as part of a wider movement incorporating humanities and social sciences and the extensions to digital humanities, e-research, and digital research infrastructures (Brügger, 2012, p. 103). Therefore, this allows for a new approach to studying a phenomenon which was not possible prior to the computational turn, the digital turn, or the invention of the web. This would imply that both digital and web historiography needs to be understood in relation to what Brügger (2018) refers to as the "digitality" of the used sources/documents as well as the methodological processes and tools used to answer the research questions (see section 2.2).

Winters (2018) further points to the need to understand "the technological contexts" in which digital media has emerged, in the same way that a

> palaeographer knows how parchment and ink were produced in the thirteenth century, and so too the digital historian will be required to know how the internet works, and how algorithms are constructed (p. 285).

Moreover, Winters (2018) notes how historians who are accustomed to working with manuscripts will also know how the manuscript was produced and came into being, "and what this means for any analysis of its content" (p. 286). Therefore, for Winters (2018) the "same is true for digital data and the social, cultural, and technical infrastructures which underpin its creation and transmission" (p. 286).

Lay (2017) emphasises how historiography is "one of the essential tools for unlocking the past" and accounts this to an understanding "that history never stands still, that it is argued over and contested." This will also hold true for digital and web historiography as new methodologies are born out of necessity to deal with the advances of the internet, web and software technologies and the continual evolution of digital media. In the meantime, other methodologies will be doomed due to software incompatibilities or obsolescence and outdated digital media formats. This is not something new. Archivists, librarians, and information professionals have been discussing it for years. The big question here is how this affects the use of digital materials for academic research, whether they are digitised, born digital on the live web or reborn digital in a web archive, and how can we ensure the use of such digital materials remain accessible, allowing for research reproducibility in the future. One answer to this is summed up succinctly by Brügger (2018) below:

> Studies of the online web have to be documented before, during, or after analysis, to provide a stable object of study and to enable peers to examine the results. Therefore, the question of archiving the web is at the core of an academic study of the web (p. 4).

Thus, we need to consider what it means to study the web by relying on both the (born digital) live web and the (reborn digital) archived web as sources for academic study.

For Brügger (2018), the web can be examined through an analytical grid of five strata: an individual web element, an individual web page, an individual website, a web sphere, the web in its entirety (p. 31). Moreover, studies of the web may have an overlap of strata cases (Brügger, 2018, p. 68). For Brügger (2018) the five strata can be applied equally to both layers of the web being "the visible/audible web in the browser, and hidden text of HTML code and associated files" (p. 31). These five strata also offer an equally applicable model for studying the archived web.

The first stratum is a web element, which Brügger (2018) describes as "any coherent semiotic entity in the form of a written element, a static image element, a moving image element, or a sound element" (p. 33). For example, this may include a banner or an image on a web page, an embedded video, a podcast, footer items, text in the form of a heading, a paragraph, or all text wrapped in a <p> tag in the HTML code. Brügger (2018) offers an example here of Morris's (2019) study of sound elements from the early web in the Wayback Machine, and Cocciolo's (2015) study, also using the Wayback Machine, to investigate whether the use of text on web pages was in decline, and if so, by how much.

The next stratum deals with a web page which Brügger (2018) describes as "whatever is presented within a single browser window" (p. 33). So, it is delimited by the "borders" of the window, but Brügger notes that while the word "page" is used it should not be understood in the same way that we think about a page of a book or a print document (Brügger, 2018, p. 34). Rather, one needs to think of it as an interactive document which contains a large element of text, but also contains images, video, audio, maps, databases, and software. Thus, for Brügger (2018) studying web pages may focus on the elements that are presented in the browser window such as text, images, and video, or by analysing "the overall composition of the web page", for the study of web design evolution for instance (p. 35). Here Brügger offers an example of Richard Rogers' study of the Google home page from 1998 to 2007 (see Rogers, 2013, 2017).

When it comes to describing the website as a stratum, Brügger (2018) proposes it to be

> an analytical unit composed of interrelated web pages [which] are connected by semantic, formal, and physical performative means, and the more consistent these three types of interrelations are, the more clearly the website will be delimited (p. 34).

In expanding further, Brügger (2018) suggests that

> what delimits the website as such is the extent to which a number of web pages, treat the same subject (semantic cohesion), resemble each other (formal cohesion), and make it possible to go from one page to another (performative cohesion) (p. 34).

Hence, the study of a website might constitute the characteristics of the web pages within the site, or the types of interrelations (Brügger, 2018, p. 34). Research which focuses on individual websites include Hofheinz's (2010) study on the history of *Allah.com* and its creators and influencers, using the Wayback Machine, the live web, and personal archiving. Also, Bødker and Brügger (2018) use a case study of *The Guardian's* website in the Wayback Machine from 1996 to 2015 to examine the shifting temporalities of online news.

The fourth stratum is a web sphere. Foot and Schneider (2006) coined the term web sphere "as a set of dynamically defined, digital resources spanning multiple Web sites deemed relevant or related to a central event, concept, or theme" (p. 20). For example, it might refer to the proliferation of interconnected web content produced because of a natural disaster or be concerned with the web content produced during a sporting event or an election campaign. Thus, it may have a temporal or geographical dimension (Rogers, 2013, p. 74; Webster, 2020, p. 2). Ogden et al. (2022) propose three types of web spheres which are of common interest for study by researchers being "(1) national web domains; (2) platforms, communities, and online ecosystems; and (3) events" (p. 7). Examples here include the work of Brügger et al. (2017) who use the Danish national web archive, Netarkivet, and the Wayback Machine to examine the development of .dk domain names and the .dk domains from 2005 to 2015. Millward (2015) uses the UK Web Archive to "build a corpus of disability websites and pages" through a web sphere of key disability organisations in the UK on the early web. Millward then tests a selection of the web material in the corpus with "code validation tools to see whether they conformed to accessibility standards as set out by the World Wide Web Consortium" (Millward, 2015). Beaudouin et al. (2019) offer another example using the French national web archive (BnF Archives de l'internet) to conduct a network analysis "of websites related to the First World War with the aim of understanding how the collective memory of the war is constructed online" (p. 440).

Finally, the fifth stratum is the web in its entirety, although for Webster (2020) "the task of studying the whole Web in terms of its content (rather than its technologies) has so far proved too vast and so has seldom been attempted" (p. 1). However, Brügger suggests that one might consider Weber's (2019) chapter on 'Browsers and Browser Wars', as an example of a study which deals with a part of the history of the whole web. Webster (2020) adds that the fifth stratum might also include "studies which illuminate the nature of the whole Web

by an analysis of elements that recur across it" (p. 1). Here, Webster (2020) offers an example of Helmond's study of the "changing purpose, use, and function of the hyperlink over time" (Helmond, 2019, p. 229). Thus, Brügger's (2018) five strata offer a useful model for combining both the (born digital) live web and the (reborn digital) archived web as sources for academic study and is a useful starting point for the scholarly use of web archives.

The value of web archives as resources for research or other purposes can also be seen through the literature. 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. Gomes and Costa (2014) offer an overview on the importance of web archives in the humanities for current and future historical research, while Healy (2019) and McTavish (2020) demonstrate the benefits of web archives for studying LGBT+ histories. 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 two 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 (Mackinnon, 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). 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 in web archives, 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 across business and law. Costa and Silva (2010) suggest that web archives provide a resource for use cases to develop company trustability profiles. Denev et al. (2011) 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 (2017b) examine the use of archived web content as evidence in a court of law.

3.3 Web Archive Creators and Users

The thesis has several overlaps with other studies which examine web archiving practices and structures, and web archive user and researcher engagement. For example, the research overlaps with studies which focus on the practices and workflows of web archiving initiatives (NDSA Content Working Group, 2012; Bailey et al. 2014; Bailey et al. 2017; Farrell et al. 2018). It also has commonalities with studies that investigate the practices of web archiving initiatives, as well as addressing the challenges for the use of web archives for research (Dougherty et al., 2010; Truman, 2016; Vlassenroot et al., 2019). Other studies which intersect with this research include user studies which focus on engagement with web archives (Jatowt et al., 2008; Costa and Silva, 2010; Moiraghi, 2018), and studies which specifically examine scholarly awareness and engagement (or non-engagement) with web archives (Hockx-Yu, 2014; Riley and Crookston, 2015; Costea, 2018). Also, worth mentioning is Truter's (2021) study which looks at research data management and sharing practices of researchers in web archive studies. The next section offers a review of a selection of these studies. To note here, the review only examines literature in English, as it is the native language of the researcher.

3.3.1 Web Archiving Practices, Tools, and Knowledge of Use

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 there is not enough 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, and their knowledge about the use of their web archive collections.

The first study was conducted in 2011 with 73 participants responding (NDSA Content Working Group, 2012, p. 11). 63 participants responded to a 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). This study also finds "an area of uncertainty" by web archiving institutions vis-à-vis how collections were being used (NDSA Content Working Group, 2012, p. 10). For example, in response to a question on how researchers are using their archive, a large majority responded with variants of "unknown"; "too soon to tell"; or "good question" (NDSA Content Working Group, 2012, p. 11). The report observes that "the lack of knowledge about web archive usage and users" is clearly a topic that merits further investigation" (NDSA Content Working Group, 2012, p. 11).

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 number of participants who responded to this question. Thus, it is difficult to get a feel for usage levels 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 question on how researchers are using web archives does not appear to have been asked in the questionnaire.

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). In response to whether the respondents had active researchers utilising their web archive, there were 80 responses of which 19% (=15) answered 'Yes', 30% (=24) answered 'No', and 51% (=41) answered 'Don't know' (Bailey, et al., 2017, p. 27). Again, the

study highlights the lack of understanding in how collections are being used as “an area of activity that merits community attention” (Bailey, et al., 2017, p. 27).

The 2017 survey (N=119) saw a majority of institutions using external services for harvesting web materials, 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 the responses 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. 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). When asked about the use of their web archives by researchers, 117 participants responded, of which 18% answered ‘Yes’, 33% answered ‘No’, and 49% answered ‘Don’t Know’ (Farrell et al., 2018, pp. 23–24). While this question specifically targets use by researchers, it provides some indications of the extent to which there is a lack of awareness by web archiving institutions apropos how their web archives are being used.

3.3.2 Web Archiving Practices, and Challenges for Web Archive Users

Sponsored by the Harvard Library, Truman (2016) conducted a study to document international web archiving programmes (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 programmes (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).

3.3.3 Web Archive User Studies

In an early study related to access and use of archived web content in Japan, Jatowt et al. (2008) conducted a survey of 1,000 internet users to gain insights on the possible types of interactions participants might have with document histories in a web archive. By document histories, they refer to the different versions of captured web pages. In examining how many participants used web archives in comparison to other web resources, it was revealed that only 1.9% of respondents used a web archive, such as the Internet Archive's Wayback Machine, at least once a month. Jatowt et al. suggest a possible reason for this, may be due to "the lack of large Web archives open to the public in Japan" and many respondents seemed to be unaware of "the existence of Web archives" (p. 11). In response to the type of information respondents would like to obtain if they could access page histories, 34.2% of participants selected the choice of information about the age of the site, and 21.1% selected the choice for information about the age of the page. In another question, related to access to past content of web pages, participants were asked what they would like to see if they could access the past content of a visited page. Participants were provided with a choice of answers, with two of the top answers being: 49.4% of participants wanted to revisit content that had already disappeared and 29.2% wanted to view content that could not previously be accessed.

Their final question concerns the types of pages for which participants would have liked to view their histories. As a first preference, 42% of participants preferred to view the histories of news sites, and 30.7% preferred to view the histories of pages related to their interests and hobbies. They find that the "types of pages for which users want to see historical data can vary from person to person. The depth to which users would like to [interact] with page histories also depends on various cases" (p. 13). Thus, they surmise that "archivists, historians or other professionals may have different requirements and needs regarding the types of documents to be archived" (p. 11). Therefore, they believe that

> end users should decide what types of data should be preserved and what types of access should be provided to gain entry to such information in order to make it popular and useful (p. 13).

They also present an interesting case for the possibility of using web archives for comparing historical information with actual information on various real-world objects, such as companies or politicians for example. Thus, they posit that "users would be entrusted with more power to assess the quality and characteristics of real-world objects, be it companies, institutions, or persons" (Jatowt et al., 2008, p. 13).

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. The study also identified several use cases which include: to collect information about a subject written in the past; to download an old file no longer available on the live web (e.g., images, software, and music); research old information like political events; and the creation of trustability profiles, based on company and employer information of the past web (Costa & Silva, 2010).

Examining scholarly engagement, Hockx-Yu (2014) offers a secondary analysis of data that was collected through a user study conducted by the British Library in 2012. The purpose of the user study was to examine the perceptions of scholars for the research value of the Open UK Web Archive, and to gather feedback on access mechanisms; identify gaps in content; and develop a better understanding of the use, or lack of use, of web archives by researchers. The article gives a comprehensive overview of the type of data that the web archive collects and how it was presented to researchers at the time of the British Library study. It further explains the challenges of balancing users' expectations alongside technical as well as legal limitations. The British Library study found that those who valued the archive the most were scholars interested in web history, statistics, and digital preservation

research. From this, Hockx-Yu summarises three approaches to engaging researchers with web archives. The first is in curating thematic collections as a research output, the second is collaborating with researchers to help them better understand what a web archive is and support them during their research project while the third is the independent use of web archives. Hockx-Yu further discusses the benefits as well as challenges with these three trends (Hockx-Yu, 2014).

On behalf of the National Library of New Zealand (NLNZ), Riley and Crookston (2015) undertook a study via a survey of academics in New Zealand in the disciplines of humanities and social sciences at seven universities and one wānanga tertiary institution (education in a Māori context). The aim of the survey was to gain some insights on the awareness of the existence of web archives by university academics, and to establish more understanding of the use of archived websites by university academics. It further intended to configure what else the NLNZ could do to assist third-level teachers in the provision of access to archived websites for educational benefits. The results and analysis were based on 257 fully completed surveys, and 33 partially completed surveys (N=290). The findings indicate a large lack of awareness by researchers in tertiary institutions in New Zealand. Other findings suggest that respondents preferred a text search option rather than the URL search, which was the only search option available for the NLNZ web archive at the time of the study. Hence, the researchers acknowledge that the access mechanism did not meet the needs of researchers. Respondents also indicated a desire for access to the NLNZ domain harvest via full text search and demonstrated a requirement for the NLNZ to develop options for making this available. Finally, 51% of researchers in the study indicated that the New Zealand Web Archive will become important for their research within the next five years (Riley & Crookston, 2015).

In Denmark, Costea (2018) conducted a study with the aims of providing some perspectives on researcher engagement with web archives, researcher needs in the use of web archives, and to identify reasons for the non-use of web archives by researchers. The study targeted professors, researchers, and PhD students from the Arts, Humanities, and Social Sciences in two Danish universities. It used a mixed method approach of an online survey (n=88), semi structured interviews (n=3) and testing with first-time users (n=2). After analysis, Costea found a noteworthy lack of awareness of the existence of web archives as resources for research. Many researchers were unaware of the content of a web archive, and how a web archive can be used as a resource for research. Users and non-users alike appreciated the value of archived web content, but also identified a need for improvements to web archives to satisfy researchers' needs in the areas of discoverability options, data selection, data

management, and more access to methods for data analysis. Issues of incompleteness of data in web archives were also mentioned. Access to more comprehensive documentation and metadata was thus seen as a requirement for researchers. Findings from both the survey and interview also highlight the need for researchers to be able to extract data from a web archive to create a dataset for their own research needs (Costea, 2018).

As part of the *Digital Library Futures* project (2017-2019), Gooding et al. (2019) conducted a study to assess the impact of UK legal deposit Non-Print Legal Deposit (NPLD) upon academic deposit libraries and their users. Their approach is "explicitly user-centric" for the exploration of "the relationship between information seeking behaviour, legal deposit institutions, and the broader regulatory and scholarly context for NPLD" (p. 12). In using the Bodleian Libraries, University of Oxford and the Cambridge University Library as case studies, Gooding et al. used a mixed methods case study approach to research two key stakeholders being UK academic deposit libraries, and users of UK academic deposit libraries. They utilised interviews, surveys, web analytics, and subject-based bibliographic analysis and examined the use of NPLD collections in the context of NPLD e-books, NPLD journals and the NPLD collections in the UK Web Archive which are only accessible onsite in one of the six UK legal deposit libraries.

From the preliminary literature, they identify five main problems. The first problem is concerned with how "NPLD in academic deposit libraries has been under-investigated" (p. 6) and note that while "academic libraries are motivated to secure access to materials for their readers [...] little has been written on how such motivations inform how academic deposit libraries approach NPLD" (p. 6). The second problem relates to the fact that there has been a minimal amount of published data regarding "the users of NPLD collections" (p. 7). The third problem is concerned with how the disciplines like digital humanities, data science, and quantitative social sciences have evolved to require "libraries to develop new forms of licencing, collection management and support for digital materials in response to user needs" (p. 7). They further point out how the UK government has supported "computational research through a 2014 copyright exception that allows non-commercial text and data mining of copyrighted materials" (p. 7). Nevertheless, they highlight how this sentiment is not extended to NPLD collections (p. 7) and suggest that the lack of planning for text and data mining is "now a significant barrier for innovative research" (p. 24).

The fourth problem is concerned with how the "NPLD regulations were introduced at a similar time to broader strategies for widening online participation" (p. 7). Here they note how government reports emphasise "inclusion and access", and how "scholarly publishing

is increasingly transitioning towards Open Access" which is also supported by government and research initiatives and highlight how "copyright regulations have been enhanced to allow the provision of accessible copies of materials for readers with a recognised disability" (p. 7). Yet, aspects such as these ,

> are not formally reflected in the NPLD regulations, which use as a basis the Copyright, Designs and Patents Act 1988 (1988) as amended by the Copyright (Visually Impaired Persons) Act 2002 (2002). This means that the 2013 regulations only allow for accessible copies of NPLD materials to be made available for readers with visual disabilities, rather than all persons with a recognised disability. As such, there is a gap in understanding of the extent to which NPLD supports emerging practices relating to Open Access and accessibility for disabled readers (Gooding et al, 2019, p. 19).

The final problem is concerned with their claim that the library sector in general lacks the use of empirical analysis for assessing the use of digital resources. Moreover, they note how the users of NPLD are "often framed as future researchers, an indeterminate and poorly defined group" (p. 8). Therefore, they posit that "there is a need to consider how approaches to evaluating NPLD can contribute to wider methodological debates in the library sector" (p. 8).

From their findings and analysis of the case studies, there are several points of interest for this research. First, they identify how library staff were "disappointed with access arrangements" due to researchers having to physically attend the library, which is "contradictory to their efforts to widen access and usage" (p. 17). However, when it came to assessing the impact of NPLD on the actual users, they "found that the libraries had not established success criteria for usage", and

> Very little user assessment had been conducted to contextualise access statistics, and internal studies had instead focused upon user experience with the NPLD user interface. However, library staff generally reported that usage of NPLD materials seemed low, and that this could largely be attributed to the access restrictions (Gooding et al., 2019, p. 18).

Therefore, they were unable to situate the user experiences in a broader framework, but nonetheless conducted a survey of "an archetypal user" of the academic deposit library.

From their findings and analysis overall, they find that while NPLD does a good job in terms of the collection of national digital heritage, it neglects the end user due to its print-centric ethos. Thus, Gooding et al. (2019) propose that the legal deposit framework should be built upon the following five tenets as outlined verbatim below.

1. The long-term beneficiaries of NPLD are users, not publishers or libraries.

2. The diversity of digital media reflect a major change in information sharing, society, libraries, and research communities, which necessitates re-evaluation of the assumption that print media remain the most useful reference point for defining access protocols.

3. Publishers are entitled to protect their commercial and legitimate interests but the impact of Open Access upon academic publishing and licensing cannot be ignored.

4. Libraries must be empowered to take actions to make collections accessible, usable, and meaningful, based on evidenced trends in user behaviour and user needs.

5. The first four tenets require continued collaboration between libraries, publishers and user groups (Gooding et al., 2019, p. 5).

Therefore, when it comes to evaluating resources like legal deposit collections, in particular the use of collections with restrictions, it needs to be clearly examined in relation to the rapidity in which technology changes the landscape for end users.

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).

3.4 Summary

This chapter examined the challenges for participation in web archive research from an international perspective (RQ2), and through the literature, it offered some insights on how to improve the conditions for conducting web archive research (RQ5). First, the chapter presented an overview of web archive research, starting with web archiving and curation, and examined some of the challenges experienced by the web archiving community. Then, it examined how the use of archived web materials for research or other purposes is less established and discussed how scholars have highlighted how academics have been slow to embrace web archives as resources for research (Webster, 2020; Rogers, 2019; Leetaru, 2019; Meyer et al., 2017; Webster, 2017b; Winters, 2017; Leetaru, 2017; Brügger, 2016; Meyer et al., 2011; Dougherty et al., 2010). Thus, it could be argued that a lack of dialogue or collaboration between the creators of web archives, and end users (or even potential end users) has had some effect on engagement with web archives for research. However, the literature also demonstrated that 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 (Winters, 2020a; Hock-Yu, 2014; Gooding et al., 2021; Vlassenroot et al, 2019; Graham, 2019; Ogden, 2021; Brügger, 2021c; Ogden, 2021; Ogden et al. 2022; Ben-David, 2021). Therefore, to understand the reasons for a lack of scholarly engagement with web archives, and the various challenges faced by this community, it is equally necessary to understand the challenges for web archiving communities and how these challenges overlap and intersect across communities of practice within web archive research.

4.0 SKILLS, TOOLS, AND KNOWLEDGE ECOLOGIES IN WEB ARCHIVE RESEARCH

The realisation of the Internet's potential to connect not just computers but individuals, families, communities and nations – through the growth of the web – has transformed our lives over the last two decades. Our histories are increasingly both created and consumed online, for an audience of millions or for an audience of only one or two people. The ease with which it is possible to write and post information online, the speed with which one can react to news and contribute to ensuing debates, has dramatically altered – in scale and type – the group of people whom we might now describe as creators, publishers or authors (Winters, 2017, p. 173).

4.1 Introduction

The previous chapter explored some of the challenges for participation in web archive research and surmised that in order to understand the reasons for a lack of scholarly engagement with web archives, and the various challenges faced by this community, it is equally necessary to understand the challenges for web archiving communities. Therefore, to move forward, there is a need for a deeper understanding of the challenges for both the creators and users of web archives, and how these challenges overlap and intersect across communities of practice within web archive research.

Through a collaborative interdisciplinary project (WARST Project), this chapter examines the challenges for participation in web archive research (RQ2) and explores ways in which to improve the conditions for conducting web archive research (RQ5). As a point of departure, the chapter considers 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, selection, capture, storage, quality assurance, preservation and maintenance, to replay/playback, access, use and reuse (Bragg & Hanna, 2013).

From there, the chapter 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 overlaps and intersections of such challenges across communities of practice. In doing so, it engages with research methods within information sciences through a survey study to collect statistical and qualitative data in the form of free text responses. The survey 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.

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 from Maynooth University Research Ethics Committee [SRESC-2021-2436150]. The research team are all members of WARCnet (warcnet.eu), and between them, have backgrounds in traditional 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.[8] 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, n.d., About WARCnet). 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, financial, 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.

In pursuit of this, the chapter 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 overlap and intersect across communities of practice, and
- explore ways to improve the conditions for conducting web archive research.

[8] WARCnet Meetings, https://cc.au.dk/en/warcnet/meetings

In the next sections, we offer an overview of related literature and discuss the methodology. We then present the findings and conclude with a discussion organised around eight main dimensions as outlined below:

- 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
- Final Thoughts

4.2 Related Literature

The research for this chapter has several overlaps with other web archive user and scholarly engagement studies (Costa & Silva, 2010; Jatowt et al., 2008; Hockx-Yu, 2014; Riley & Crookston, 2015; Costea, 2018; Moiraghi, 2018). However, the chapter also 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. Therefore, the research for this chapter has some overlaps 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). There are also commonalities with the work of Dougherty et al. (2010), Truman (2016) and Vlassenroot et al. (2019) who investigate the practices 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. A review of some of this literature is available in chapter 3.0.

4.3 Methodology

In this section, we lay out the methodological approach for the chapter, which includes the survey design, and approaches for data collection and analysis. The research for this chapter was conducted in compliance with best practice guidelines for the collection and management of research data, as outlined in Maynooth University Research Ethics Policy (2019), Maynooth University Research Integrity Policy (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 thesis 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, 2021).

4.3.1 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. The 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 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. Permissions were also sought from participants for the publication of 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 six 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.

4.3.2 Survey Software

We utilised the JISC Online Surveys tool for collection purposes (Joint Information Systems Committee). Maynooth University provides, to staff and PhD students, access to this software for academic and research purposes. To note here, it is currently the only tool permitted by the university for conducting online survey studies of this nature.[9]

[9] The use of the tool is subject to the terms and conditions set forth in Maynooth University Online Surveys User Policy (2019) 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.

4.3.3 Survey Recruitment

The focus of the chapter 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 et al. (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.

4.3.4 Survey Responses

The survey was open from 21 July to 23 September 2021. We anticipated 25 to 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 instances. Therefore, the final tally of complete surveys for analysis is 44 respondents. In a Danish study on scholarly awareness and engagement with web archives, Costea (2018) 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" (p. 11). 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).

4.3.5 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.[10] 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.

4.3.6 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

[10] IIPC, Tools & software, https://netpreserve.org/web-archiving/tools-and-software/; NetLab, Tools and Tutorials, https://www.netlab.dk/services/tools-and-tutorials/

encourage completion (Chudoba, 2018; CoolTool, 2017; Steber, 2016). As mentioned previously in section 4.3.3, while the focus of the survey 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, as discussed in section 4.4.1.2 (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 survey cannot be construed to represent any target group population as a whole.

4.4 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=).

4.4.1 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.

4.4.1.1 Participant age and gender

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

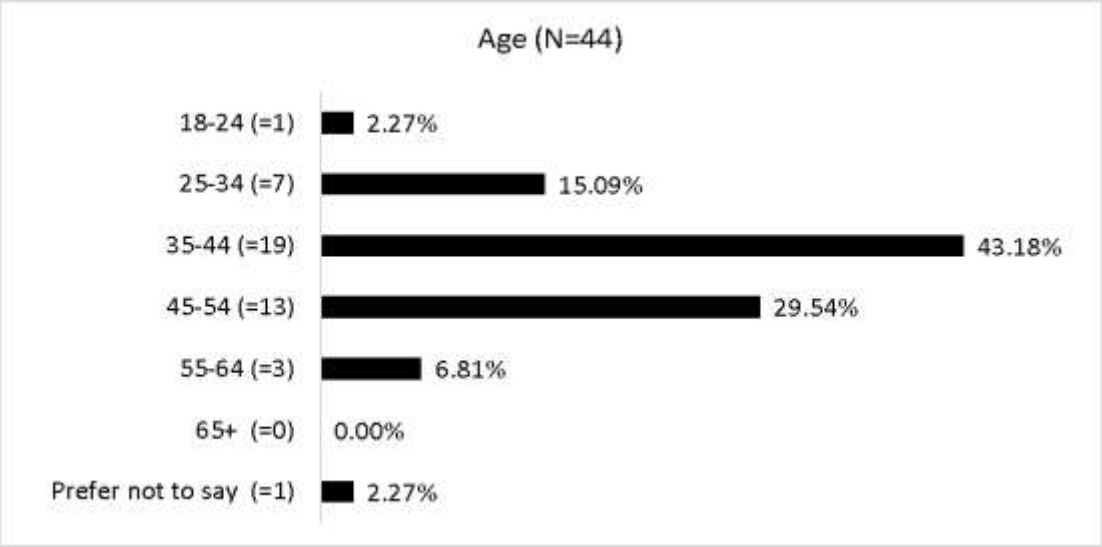

Figure 4.1: Representation of participant responses for age (N=44)

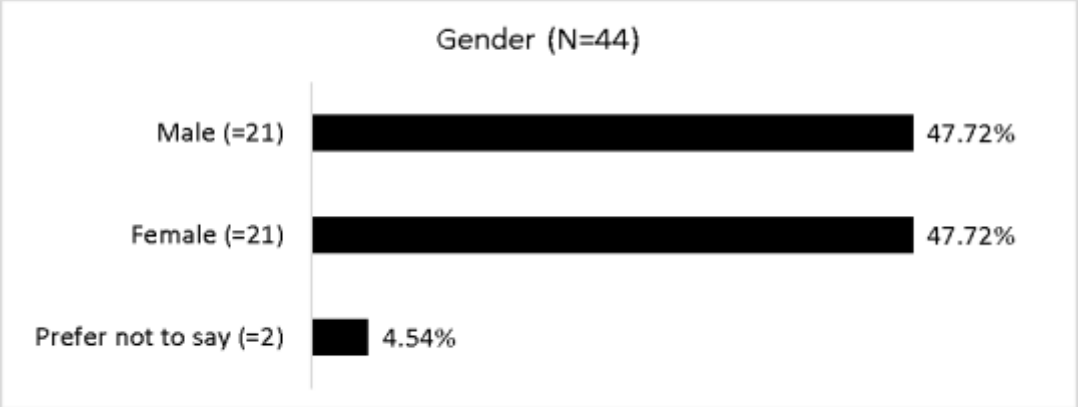

Figure 4.2: Representation of participant responses for gender (N=44)

4.4.1.2 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.

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

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.

4.4.1.3 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 practices 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) ● Web archives, web archiving, curation (=25) ○ Foster pathways for research access/use (r=14) ○ Collection development/strategies (r=4) ○ Web archiving/curation (in general) (=4) ○ Web archiving and metadata (r=2) ○ Web archives - compliancy for linked open data standards (r=1) ● Archives and records management (r=8) ● Digital preservation, long-term preservation (r=6) ● Libraries and digital libraries (r=7) ● Digital preservation, long-term preservation (=6) ● Documentation (institutional/organisational) (r=2)	r=55

● Media formats (r=2) ● Email archiving (r=1) ● Information literacy (r=1) ● Literature evolution (r=1) ● Museum studies (r=1) ● Open access and scholarly publication (r=1)	
|> Arts, Humanities, DH, Social Sciences, Media Studies ● History (r=10) ● Culture and heritage (r=5) ● Languages, Linguistics, Semiotics (r=4) ● Identity and Memory (r=3) ● Anthropology (r=1) ● Archaeology (r=1) ● Cinema (r=1) ● Egyptology (r=1) ● Ethnography (r=1) ● Politics (r=1) ● Psychology (r=1) ● Sociology (r=1)	r=30
|> Internet/web applications, systems, histories ● Web design/ designers (r=2) ● Privacy and consent online (r=1) ● Vernacular web (r=1) ● Web based information systems (r=1) ● Web based learning (r=1) ● Web tracking (r=1)	r=7
|> IT/Computer applications, systems, environments ● User experience (UX) design (r=2) ● Artificial intelligence (r=1) ● Information technology (r=1) ● IT system architecture (r=1) ● Text recognition (r=1)	r=6
|> Research practices and approaches ● r: “archived web as a source” ● r: “evolving research practices with born digital material“ ● r: “The impact of changing technology on historical research practice.” ● r: “Longitudinal in nature - both from a DH perspective and a technical one.” ● r: “digital methods for humanities research”	r=5

|> Audiovisuals, Music, Video Games	r=4
|> Design related interests ● Design & Anthropology (r=1) ● Design education (r=1) ● Design history (r=1) ● Design pedagogy (r=1)	r=4
|> Law ● Case law (r=1) ● Regulations (r=1) ● Legislation (r=1)	r=3
|> Transnationalism, Migration	r=2
|> Reading	r=1
|> Travel	r=1
|> In-vivo representations ● r: “Probably broader than they should be!“	r=1

4.4.2 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.

4.4.2.1 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

4.4.2.2 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 4.4.1.2. 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=).

4.4.2.2.1 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 ● Browser-based crawlers (r=23) ○ Conifer (prior, Webrecorder) (r=9) ○ ArchiveWeb.page (r=4) ○ Brozzler (r=4) ○ Electrolyte (r=3) ○ Browsertrix (r=2) ○ Umbra (r=1) ● Crawl software in general, not browser-based (r=13) ○ Heritrix (r=11) ○ HTTrack Website Copier (r=1) ○ Wget (r=1) ● Web crawler (in general) (r=1)	r=37
|> Curating web archive collections: selection, configuring and scheduling crawls, annotating seeds, performing QA ● NetarchiveSuite (r=5) ● CWeb (r=2) ● W3ACT (r=1) ● Web Curator Tool (r=1) ● r: "selecting material for collection"	r=10
|> Accessing/replaying archived web data ● Internet Archive, Wayback machine (r=3) ● OpenWayback (r=2) ● pywb (r=2) ● waybackpy (r=1)	r=8
|> Managing data ● Excel, spreadsheet, .csv (r=3) ● CMS, Cloud platforms (r=2) ○ DSpace (r=1) ○ Google Drive (r=1)	r=5
|> Finding source material (r=4) ● Internet, search engines, web search (r=2) ● Library catalogues and databases (r=2)	r=4
|> Tools with diverse purposes (=4)	r=4

● Browser tools (r=1) ● command-line tools (r=1) ● Python scripts/libraries (r=1) ● r: "the type of tools that come for standard with a PC"	
|> Collecting data from API ● Instaloader (r=1) ● Social Feed Manager (r=1)	r=2
|> Screenshot, screen capture ● screen capture tools (in general) (r=1) ● snipping tools (in general) (r=1)	r=2
|> Digital forensics/preservation ● MediaArea tools (r=1)	r=1
|> Web archiving subscription services ● Archive-It (r=1)	r=1
|> In-vivo representations ● r: "In house developed web archiving tools"” ● r: "institutional sources" ● r: " text recognition evaluation tools"	r=3

4.4.2.2.2 Scholar, academic, lecturer, student, or IT/web design environment

Table 3.5 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 ● Browser-based crawlers (r=3) ○ Conifer (prior, Webrecorder) (r=2) ○ Browsertrix (r=1) ● Crawl software in general, not browser-based (r=4) ○ Heritrix (r=2) ○ HTTrack Website Copier (r=1) ○ Wget (r=1)	r=7
|> Finding source material ● In libraries/web archives (r=3) ○ SHINE tools - UKWA (r=2) ○ Library catalogues and databases (r=1) ● Internet, search engines, web search (r=3) ○ Internet (r=1) ○ Search engines / web search (r=2)	r=6
|> Screenshot, screen capture, screencast ● screenshot tools/functions (in general) (r=2) ● script for screenshot automation (r=1) ● Snagit (r=1) ● Websnapper (r=1)	r=5
|> Tools with diverse purposes ● Browser tools (r=2) ● Python scripts/libraries (r=1) ● R (Rstudio) (r=1)	r=4
|> Manual/scripted file downloads	r=3

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

4.4.2.3 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 4.4.1.2. 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=).

4.4.2.3.1 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 ● CDX queries/files (r=2) ● SolrWayback (r=2) ● SQL (r=2) ● Amazon Athena (AWS) (r=1) ● Apache Solr (r=1) ● ElasticSearch (r=1) ● HeidiSQL/MariaDB (r=1) ● Apache Lucene (r=1) ● NutchWax (r=1) ● r: "Web Archive user interface, faceted functions"	r=13
|> Data extraction, cleaning, transformation ● Excel, spreadsheets (r=5) ● Archives Unleashed Toolkit (r=1)	r=6
|> Programming/scripting languages, computing environments ● Python/Python libraries (r=3) ● Command-line tools (r=1) ● Jupyter Notebooks (r=1) ● R (r=1)	r=6
|> Visualisation ● Tableau (r=2) ● Kibana (r=2)	r=4
|> Digital forensics/preservation ● DROID (r=1) ● BitCurator (r=1) ● MediaArea tools (r=1)	r=3
|> Distributed processing ● Apache Hadoop (r=2) ● Apache Spark (r=1)	r=3
|> Metadata, crawl logs ● Crawl logs (r=2) ● r: "Metadata"	r=3

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

4.4.2.3.2 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 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 ● Excel, spreadsheets (r=4) ● Archives Unleashed Cloud (r=1) ● Archives Unleashed Toolkit (r=1) ● OpenRefine (r=1) ● Pattern matching (r=1) ● Regular expressions (r=1)	r=8
|> Programming, scripting languages and computing environments ● Bash/shell scripting languages (r=3) ● Python/Python libraries (r=2) ● Command-line tools (r=1) ● Perl (r=1) ● R (r=1)	r=8

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

4.4.2.4 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 ● Spreadsheets (r=16) ● .csv files (r=2) ● Excel (r=1)	r=19
|> Screenshots	r=13
|> Text related ● Text fragments/extracts (r=7) ● Quotes (r=2) ● Text (r=2)	r=11
|> Visualisations ● Graphs (r=5)	r=10

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

|> Multi-media outputs ● Twitter tweets (r=1) ● Wiki content (r=1)	r=2
|> Statistics	r=2
|> APIs	r=1
|> Digital forensics/preservation ● r: "Reports from BitCurator"	r=1
|> Evidence collection ● 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 ● r: "Image/textual search services online" ● r: "structured corpora" ● 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

4.4.3 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.

4.4.3.1 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 4.4.1.3, 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) ● Web archives, web archiving, curation (r=29) ○ Collection development (r=5) ○ Crawling (r=3) ○ Preservation (r=3) ○ Quality assurance (r=3) ○ Web archiving (in general) (r=3) ○ Curatorial management (r=2) ○ Promoting use of web archives for research (r=3) ○ Comparing transnational collection/curatorial processes (r=1) ○ Curating web archive collections for research (r=1) ○ Evaluating archival rate of national websites (r=1) ○ Information retrieval (r=1) ○ Metadata (r=1) ○ Social media archiving (r=1) ○ Web archive solutions (r=1) ○ Web archiving, history/evolution (r=1) ● Documentation & publications (r=5) ● Archival studies (r=2) ● Libraries and social media communications	r=38
|> Arts, Humanities, DH, Social Sciences, Media Studies ● Internet and web histories (r=7) ○ r: "internet literature history" ○ r: "Historical studies of the development of the [...] web" ○ r: "History of the [national] internet" ○ r: "history of websites (and the user experience of that) at the web archives"	r=23

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

4.4.3.2 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 3.4.1.2. 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=).

4.4.3.2.1 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 ● Web archivist/curator - job related (r=11) ● Promote/support research engagement with web archives (r=4) ● Institutional need (r=2) ● Digital legal deposit (r=1) ● Promote inclusive archiving (r=1) ● Promote value of web archives to stakeholders/funders (r=1) ● r: "It is the present and future of archival work." ● r: "A specific collection for a current [...] senator requires capturing his current website" ● 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 ● Preserve documentary heritage (r=1) ● 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." ● r: "Loss of content as websites/databases are updated/retired/allowed to fail"	r=3
|> Interests in research aspects/outputs of collections ● 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.” ● 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 ● 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." ● r: "old websites as primary sources from about a decade ago"	r=2
|> Business need for a law firm library ● 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 ● r: "Availability during pandemic" ● r: "An adviser taught me how to use it." ● r: "Internet Archive's Wayback Machine was an early fascination of mine." ● r: "My PhD Thesis"	r=4

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 ● Resource for historical research (r=3) ● Resource for studying migrants/migration (r=2) ● Resource for research of evolution of web design (r=1) ● Resource for research of educational broadcasting (r=1) ● Resource for internet studies research (r=1) ● r: "authoritative source" for research ● 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 ● Website obsolescence (r=1) ● Preservation for the future (r=1)	r=2
|> Ease of access to public web archives ● r: "Having ready access to web archives, which coincided with emerging research questions" ● r: "Ease of access"	r=2
|> Resource to find information/literature ● r: "Wanting to find data" ● r: "online literary magazine which is not live again but important evidences in [...] literary history"	r=2
|> Business need ● Web content strategy ○ 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 ● r: "Fascination with the centrality of the web in everyday lives and yet its propensity to obsolescence and research oversight" ● r: "Wanting [to] make data available" ● r: "Web archiving is [a] very important topic, which is not researched enough"	r=3

4.4.3.3 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 (4.54%, 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)

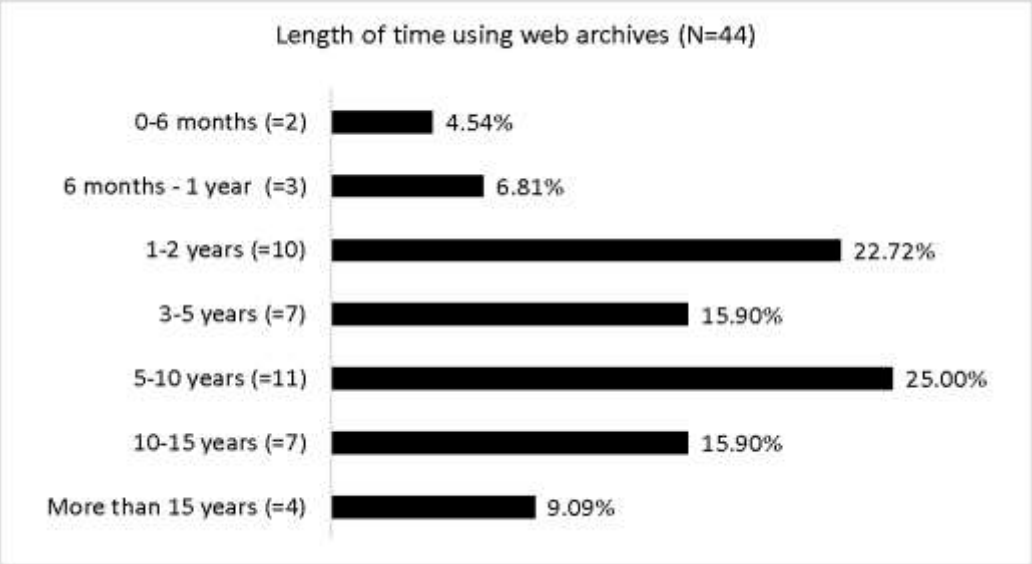

Figure 4.3: Representation of participant responses for the length of time using web archives (N=44)

4.4.3.4 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.

- 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

4.4.3.5 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 4.4.1.2. 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=).

4.4.3.5.1 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), broken links to files (r=1), and the volume of data (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 ● Broken links to files (e.g. PDFs, Excel etc.) (r=3) ● Erroneous crawls (r=3) ○ r: "Incomplete or erroneous crawls" ○ r: "The harvest is not always totally fine" ○ r: "when it gives errors in the capture" ● Layout/visual deficiencies (r=2) ○ r: "Sometimes the images are blurred" ○ r: "the visualization is not always right" ● Capturing dynamic content (r=1) ○ r: "the shallow delivery of dynamic content due to the limitations of the bots." ● Inconsistency with crawl frequency of early websites (r=1) ● r: "Variation in what is collected over time" (r=1)	r=11
|> Technical challenges ● Challenges to save sites due to firewall/security (r=1) ● Data storage (r=1) ● Data processing (r=1) ○ 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." ● Difficult to create bulk data sets/share with researchers (r=1) ● File format obsolescence (r=1) ● Lack of IT infrastructure (r=1) ● Search and discovery challenges (r=1) ● Technical challenges (in general) (r=1)	r=8
|> Legalities for acquisition/providing access ● Challenges to provide access due to legislation, copyright and GDPR (r=5) ● Acquisition challenges for selective archiving (r=2) ○ Challenges to get permissions (r=1) ○ Acquisition restrictions for selective archiving (r=1) ● Embargoes (r=1)	r=8

|> Challenges with learning new skills ● r: "complexity of the WARC files" ● 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" ● r: "Learning how to use research tools (from a non-technical user's perspective)" ● r: "Limited technical skills to analyse the WARC-files and the information within them" ● r: "learning curve was steep" ● 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 ● Cost of storage (r=1) ● Cost of services (r=1) ● Attaining funding (r=1) ● r: "On-premises access to web archives makes them economically inaccessible."	r=4
|> Documentation/metadata ● r: “confusing records” ● r: “Trying to guess the date when the site may have been crawled and when changes happen”	r=2
|> Volume of data ● r: “The size of the collections and the difficulty of narrowing down a set of data that is manageable and appropriate” ● r: “scale of the archive”	r=2
|> Conceptual challenges	r=1
|> Institutional challenges ● 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 ● r: “Having access to the raw data, as a web archivist, is very beneficial“	r=1

4.4.3.5.2 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 ● Inconsistent in terms of what was saved (r=6) ○ r: “in terms of content: sometimes the website or the entry I am looking for is not archived” ○ r: “Many websites are hardly accessible, not enough material saved.” ○ r: "Missing image files” ○ r: "Broken links” ○ r: "inconsistent in terms of what was saved” ○ r: "inaccessible website” ● Inconsistent temporal coverage (r=2) ○ r: “Incomplete temporal coverage“ ○ r: “inconsistent in terms of what was saved and when” ● Layout/visual deficiencies (r=1) ○ r: “Incorrect layout (in relation to live web)” ○ r: “Incompleteness in the data itself”	r=10
|> Legalities on access, use, and storage ● Legal challenges regarding access to data (r=4) ● Legal challenges regarding use of data (r=2) ● Inability to download data (r=1) ● Legal challenges regarding storage of data (r=1)	r=8
|> Challenges with learning new skills ● Having to acquire new programming skills (=2) ● Challenges with tools for web archive research (=1) ● Learning about the limitations of replay interfaces (=1) ● Difficulties to understand how web archives are set up (=1) ● Learning what a WARC file was (=1)	r=7
|> Research methods and approaches ● Lack of research methods/theory (r=2) ● 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: “[research] community doesn't have enough [epistemological] assessment of web archives as historical sources yet. And this is crucial for interpretation.”	r=5

● Archived web as a source for research (r=1) ● 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" ● Combining traditional methods with web archive research (r=1) ● 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." ● Data analysis (r=1) ● r: "limited analytic functionality in web- based access interfaces"	
|> Challenges in an IT/Business/Administrative environment ● r: "Funding and low awareness from stakeholders" ● 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 ● r: "issues relating to the lack of documentation" ● r: "lack of archival context"	r=2
|> Volume of data for research ● r: "volume " ● r: "Working with large-scale data"	r=2
|> Performance related issues	r=1
|> In-vivo representations ● r: "One of the big barriers was getting started" ● r: "once I wanted to get more involved, who to contact!" ● r: "Too many to count!"	r=3

4.4.3.6 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 three 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

4.4.3.7 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 ● Analytical thinking (r=2) ● Historical research skills/methods (r=2) ● Archival research skills (r=1) ● Digital humanities skills/methods (r=1) ● Mathematics (r=1) ● Understanding of provenance (r=1)	r=8
|> Information sciences (information studies) ● Archiving PDF/Screenshot, type of web archiving (r=1) ● Data management skills (r=1) ● Document database management systems (r=1) ● Library information science (r=1) ● Media formats (r=1) ● Preserving net art (r=1) ● Records management (r=1) ● Semantic web technologies for digital libraries (r=1)	r=8
|> Programming, scripting languages ● Programming tools (in general) (r=2) ● JavaScript (r=1) ● Perl (r=1) ● PHP (r=1) ● Unix shell (r=1)	r=6
|> Data analysis skills ● Visual / multimodal analysis skills (r=2) ● Pre-processing data (r=1) ● Semiotic analysis skills (r=1)	r=4
|> Website design/browser developer tools ● Browser developer tools (r=1) ○ r: "optimizing use of browsers' dev tools" ● Website design (r=3) ○ Web design (in general) (r=1) ○ r: "Looking at websites as objects (some static, some changing) helped in grasping web archiving conceptually." ○ r: "a background creating flash and CSS websites"	r=4

|> Finding information/services ● r: “trying different keywords, URLs, thinking about the way information in an organization might be organized.” ● r: “some training in finding things in libraries”	r=3
|> Software and tools ● Maths tools (r=1) ● MySQL (r=1) ● 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

4.4.3.8. 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 ● Data extraction, cleaning, and management (r=3) ○ Data cleaning tools (r=1) ○ Excel (or other spreadsheet) (r=1) ○ Regular expressions/Regex (r=1) ● Distributed processing (r=2) ○ Hadoop (r=1) ○ Spark (r=1) ● Computing infrastructure (r=1) ○ Amazon Web Services (r=1) ● Crawling software (r=1) ○ Heritrix: basic-advanced profile knowledge for functionalities (r=1)	r=7
|> Web/internet related skills ● Web design/development (r=4) ○ Web design/development tools (=1) ○ Understanding of HTML (r=1) ○ r: "Understanding how websites have been built over the past 30+ years." ○ r: "How websites are built/ made/ updated" ● Better understanding of the technical history of the web (r=1) ● Better understanding of technical history of the internet (r=1) ● How the internet works (r=1)	r=7
|> Programming, scripting languages ● Programming (r=2) ○ Programming (in general) (r=1) ○ 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." ● R (r=2) ● Python (r=1)	r=5

|> Finding information/services ● r: “A list of more web archives” ● r: “topical knowledge about where to look”	r=2
|> Application of metadata ● r: “Information on how best to assign metadata”	r=1
|> Collaborative skills ● r: “How to collaborate with others”	r=1
|> Digital legal deposit ● r: “How digital legal deposit works and what it is”	r=1
|> Ethnography	r=1
|> Glossary of terminology ● r: “A glossary of terminology would also be helpful”	r=1
|> Managing protected data ● r: “Handling protected data (sensitive data and copyright protected data)”	r=1
|> Marketing and public relations	r=1
|> In-vivo representations ● r: “how indexes are generated, what they contain, and the potential uses they can be put to“ ● r: “(hyper)link tracing / retrieval would be useful” ● r: “I really use web archives in a limited capacity and I am not trying to get too fancy.” ● r: “All the necessary skills were provided by the [web archive] team” ● r: “Sustainability (long-term availability) of the Internet Archive’s Wayback Machine”	r=5

4.4.3.9 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 the thematic representations 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 ● How web archiving works (r=17) ○ Understanding of web archiving tools (r=4) ○ Web archiving (in general) (r=3) ○ How crawling/capture works (r=2) ○ Understanding of data storage (r=2) ○ Understanding of playback/replay (r=2) ○ Understanding of WARCs (r=2) ○ How to create web archiving workflows (r=1) ○ How web archives are organised (r=1) ● Educational activities for web archiving (r=1) ● International collaboration on web archiving (r=1) ● Web archiving standards (r=1) ● Other representation (r=1) ○ r: "Implementing foreign professional concepts into our own web archiving practice."	r=21

|> Software and tools ● Data extraction, cleaning, and management (r=5) ○ Excel, spreadsheets (r=3) ○ Regex/ Regular expressions (r=1) ○ r: "Tools for data cleaning" ● Crawling software (r=2) ○ Heritrix (r=2) ● Network analysis (r=3) ○ Gephi (r=3) ● Curating collections: selection, configuring and scheduling crawls, annotating seeds, performing QA (r=2) ○ CWeb (r=1) ○ NetArchiveSuite (r=1) ● Distributed processing (r=2) ○ Hadoop (r=1) ○ Spark (r=1) ● Replaying archived web data (r=1) ○ Open Wayback (r=1) ● Web archive access and analysis ○ GLAM Workbench (Jupyter Notebooks) (r=1) ● Computing infrastructure (r=1) ○ Amazon Web Services (AWS) (r=1) ● r: "using dev tools"	r=18
|> Digital curation processes/workflows ● Metadata (r=6) ● Long-term preservation/infrastructures (r=3) ● Access (r=2) ● Collection (r=2) ● Digital storage (r=2) ● How digital curation works (r=2)	r=17
|> Data analysis skills ● Data analysis (in general) (r=3) ● Link analysis (r=1) ● Quantitative data analysis (r=1) ● Qualitative data analysis (r=1) ● Text analysis (r=1) ● Visual analysis (r=1) ● Large-scale data analysis (r=1) ○ r: “Understanding better the challenges and potential for large-scale data analysis.”	r=9
|> Programming/scripting languages ● Programming and visualisations with R (r=4)	r=7

● Python scripts/libraries (r=2) ● Shell scripting (r=1)	
|> Web/internet related skills ● r: “Above all, how the creation of the Web works and behaves in general” ● r: “How websites are updated” ● r: “How the internet works - Geo-IP, servers, browsers, domains, hosting etc. “	r=3
|> Research methods and approaches ● r: “Knowing more about research uses of archived web” ● r: “theoretical approaches to web archives and source code.” ● r: “how to keep notes about where information/data comes from”	r=3
|> Database creation and maintenance	r=1
|> Digital legal deposit ● r: “How digital legal deposit works and what it is”	r=1
|> Fair use, copyright, reproduction rights ● r: “How Fair Use works - copyright, reproduction rights, fair use”	r=1
|> Managing protected data ● r: “Handling protected data”	r=1
|> In-vivo responses ● 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.” ● r: “Most of my digital skills!”	r=2

4.4.3.10 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)

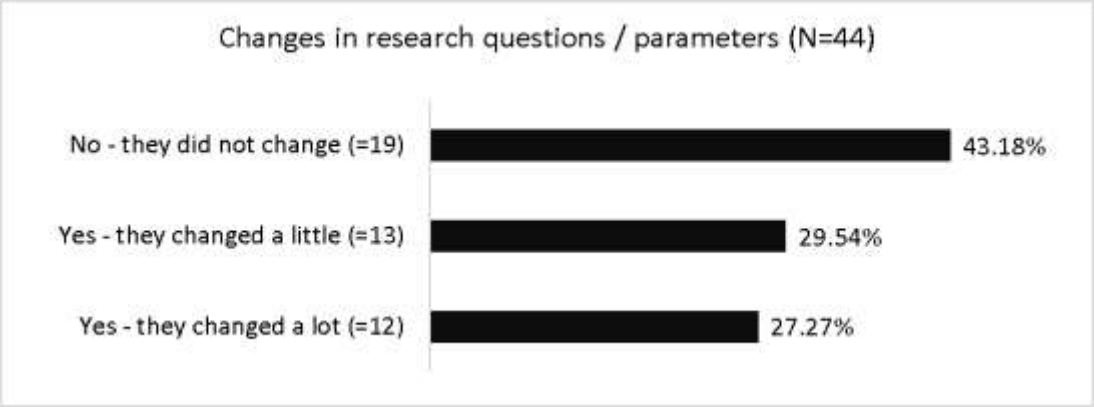

Figure 4.4: Representation of participant responses for changes in research questions or parameters (N=44)

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 that 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)

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 ● Data analysis, data cleaning (r=4) ○ 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“ ○ 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.” ○ r: “The opportunities and tools available for large-scale data analysis has changed quickly during the time I have worked with web archives” ○ r: “I always find that digging into some data gives me new ideas for new things I can dig out.” ● Access to raw data (r=1) ○ 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” ● Attendance of online conferences/webinars (r=1) ○ 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.” ● Blog design/communities (r=1) ○ 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.”	r=11

● COVID disruption (r=1) ○ r: “Covid has disrupted travelling to individual libraries to consult datasets.” ● Data centred approach (r=1) ○ r: “Completely new data centered approach” ● Digital humanities tools/methods/approaches (r=1) ○ r: “Digital Humanities and using large scale computation methods and tools like Hadoop/Spark through R with Jupyter Notebooks and other similar tools“ ● Statistical analysis requirements (r=1) ○ r: “The requirement for better knowledge of using spreadsheets in [statistical] analysis “	
|> Web archives, web archiving, curation ● Collection development strategies/decisions (r=4) ○ r: "collaborative archiving” ○ 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.” ○ 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: “At the beginning, more administrative-type pages were collected, later it was expanded to more cultural topics.“ ● Challenges with social media archiving (r=2) ● Learning automation processes (r=1) ● Priorities change (r=1) ○ r: “times have expanded and interest was no longer a priority”	r=8
|> In-vivo representations ● r: “I find that with every project, the more you learn, the more you refine the question and the parameters for the search.” ● r: “I'm a reference librarian, so my research projects are always changing.” ● r: “It has been a process of constant development, since I have not been bound into a clearly bounded project as such.” ● r: “looking at specific types of written sources” ● 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

4.4.4 Citation Practices

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

4.4.4.1 Referencing styles and practices

Participants were asked about the referencing systems they use for citing sources in their research in general, and 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)

In addition, 22 participants entered free text responses for 'Other' referencing systems. The responses 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 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)

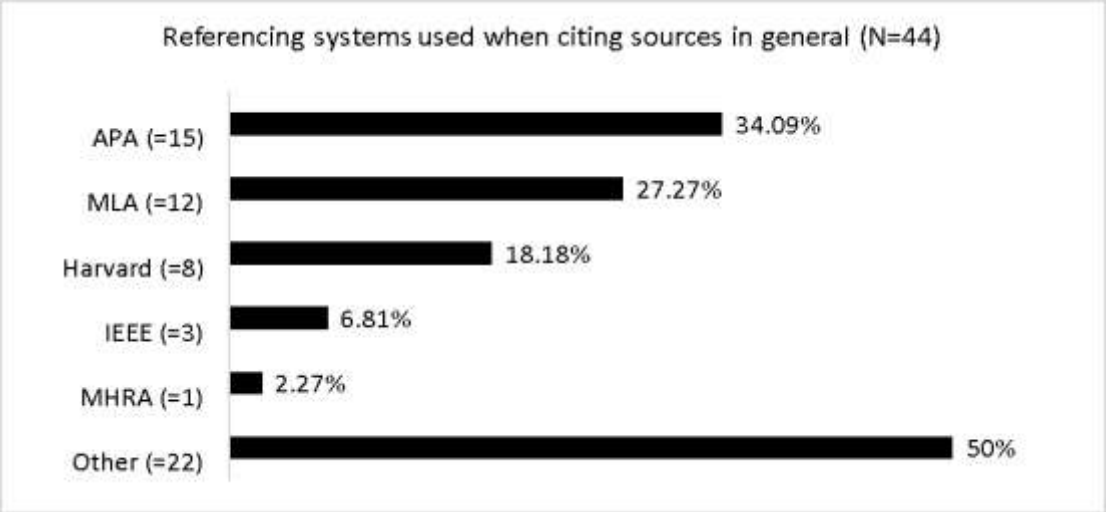

Figure 4.5: Representation of participant responses for referencing systems used when citing sources in general (N=44)

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 ● Chicago (r=6) ● Turabian (r=1)	r=7
|> Other standards/specifications ● ISO standards (r=2) ● Digital Object Identifier (r=1) ● r: "Use DOIs to cite datasets where they exist. (e.g. UK Web Archive derived datasets)" ● ISBD (International Standard Bibliographic Description)(r=1) ● RDA (Resource Description and Access) (r=1) ● ГОСТ (GOST) (r=1)	r=6
|> Non-applicable for some participants	r=4
|> Depends on journal/publisher/proceedings	r=2
|> Internal/institutional citation formats ● r: "Tend to use an internal format" ● r: "Our institutional citation formats are unique and varied"	r=2

|> Reference management applications/mark-up (for any style) ● Zotero (r=1) ● LaTeX/BibTex (r=1)	r=2
|> In-vivo representation ● r: "I haven't written academic papers citing web archives (generally, I write policy papers that are about web archiving)" ● r: "Not yet published"	r=2

4.4.4.2 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 participant's position and indicates that there is no relevant pattern or differentiation between one community of practice or the other.

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) ● Sometimes (n=11) ● Yes (n=4)	● No (n=8) ● Sometimes (n=5) ● Yes (n=1)

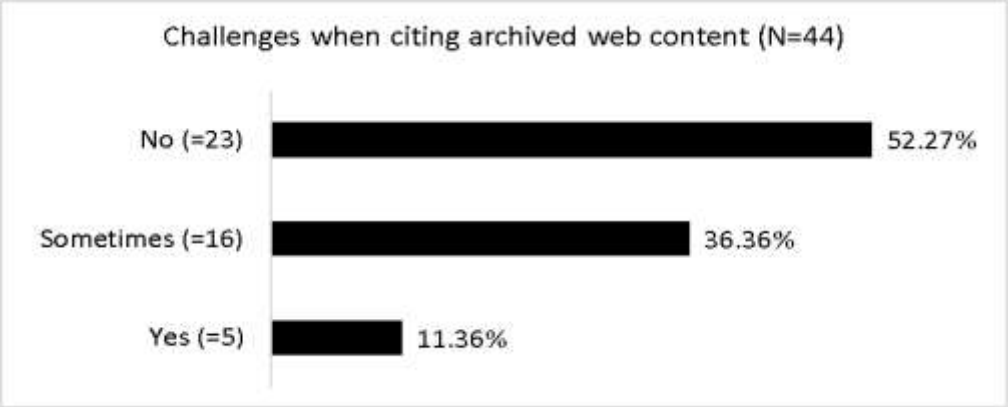

Figure 4.6: Representation of participant responses for challenges when citing archived web content (N=44)

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 ● Lack of guidelines (in general) (r= 1) ● r: "Agreeing on best practice" ● r: "Lack of rules for citing 'popular' things like forums (or more recently, but less likely to be archives, social media)" ● r: "Sometimes it is not quite clear what the best way to cite a source is." ● r: "Referencing standards are sometimes not adapted to the archival materials." ● r: "The existing systems don't have a model for this type of content." ● 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 ● r: "Citing historic content in a closed archive only accessible by other researchers in a [persistent] way ● r: "Copying and pasting a URL from a reading room viewer is not possible as the browsers are locked down." ● 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)." ● 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."	r=4
|> Uncertainties for citing archived web content ● Should it be cited like a normal website? (r=1) ○ r: "It is difficult to know if you should cite it similar to a website" ● Should the source be treated as a normal URL? (r=1) ○ r: "Unsure whether to treat it is a URL" ● Should the web archive be acknowledged? (r=1) ○ r: "whether the archive should be acknowledged" ● What dates should be used? (r=1)	r=4

○ r: "what dates should be used (capture date, access date, date of original publication, e.g. a blog post or article)."	
|> Challenges specific to the URL for archived web content ● r: "The standard URL identifier derived from Wayback, while adequate, is unwieldy and not easily read by humans." ● r: "Ensuring stability of references, even if archive systems change"	r=2
|> Not easy to cite sources from a web archive (in general) ● r: "Web Archives tend not to offer an easy way to generate a citation." ● 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 ● 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 ● r: "The web address is not stable" ● 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: "References can be either incomplete, not cited correctly or incorrect which requires further research." ● 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."	r=4

4.4.4.3 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)

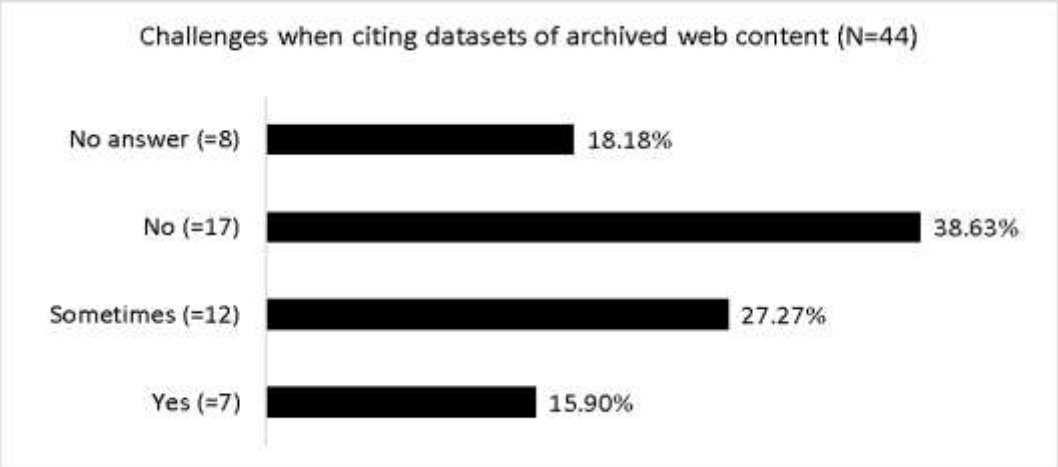

Figure 4.7: Representation of participant responses for citation challenges using datasets of archived web content (N=44)

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)

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 ● r: "It's just hard to reference, there are almost no guidelines on the subject." ● r: "I don't know if there is an agreed standard for citing datasets." ● r: "Sometimes it is not quite clear what the best way to cite a source is." ● r: "The existing systems don't have a model for this type of content" ● r: "Referencing standards are sometimes not adapted to the archival materials"	r=5
|> Amount of data / details to include in a dataset citation ● 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."	r=3
|> Not easy to cite datasets (in general) ● Not easy to cite datasets (r=2) ● 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 ● Should the web archive be acknowledged? (r=1) ● What dates should be used? (r=1)	r=2
|> Data/content reliability within a dataset ● 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	r=1

● r: “Derived data sets from web archived data may not be properly preserved”	
|> System restrictions ● r: “we don't always have ways of recording the source of web content in our systems.”	r=1
|> In-vivo representations ● r: “Unable to recall” ● r: “It is not a task that I do continuously”	r=2

4.4.5 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.

4.4.5.1 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 ● International Internet Preservation Consortium (r=5) ○ r: "IIPC Congress and workshops about tools" ○ r: "IIPC webinars, workshops" ○ r: "Training course from the IIPC" ○ r: "IPC sponsored events" ○ r: "IIPC Webinar about Web Archive" ● Training from a web archive (r=3) ● Archives Unleashed Datathons (r=2) ● Institutional training/courses (r=2) ● Online training/tutoring (r=2) ● RESAW (r=2) ○ r: "workshop at RESAW conferences/meeting"	r=26

○ 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." ● Training/courses (in general) (r=2) ● Workshops (in general) (r=2) ● MODE Summer School, UCL, Institute of Education, Knowledge (r=1) ○ r: "Multimodality Summer School (week-long at Institute of Education / Knowledge Lab)" ● Netlab, Aarhus University (r=1) ○ r: "Netlab - course by Aarhus University" ● Rhizome (r=1) ○ "[lecture] by Dragan Espenshied from Rhizome" ● The National Archives UK/ Digital Preservation Coalition (r=1) ○ r: "Novice to Knowhow from TNA and DPC" ● Training from a digital repository (r=1) ● Trinity College Dublin (r=1) ○ r: "Digital Humanities course run by Trinity College Dublin"	
|> Software and tools ● Data analysis, cleaning, transformation (r=6) ○ Archives Unleashed Toolkit (r=2) ○ Excel (advanced) (r=1) ○ Pandas (r=1) ○ Power BI (r=1) ○ Tableau (r=1) ● Crawling software (r=3) ○ ArchiveWeb.page (r=1) ○ Conifer (prior, Webrecorder) (r=1) ○ Heritrix (r=1) ● Network analysis (r=3) ○ Gephi (r=2) ○ LinkGate (r=1) ● Information retrieval (r=2) ○ Solrwayback (r=2) ● Programming, scripting languages and computing environments (r=1) ○ Jupyter Notebooks (r=1) ● Web archive access and analysis (r=1) ○ GLAM Workbench (r=1)	r=16
|> Websites, web pages, blogs ● International Internet Preservation Consortium (r=7) ● Zenodo (r=2) ● ArchiveWeb.page (r=1) ● Conifer (r=1) ● Heritrix (r=1) ● One Terabyte of a Kilobyte Age (Blog) (r=1) ● Pandas (r=1) ● SolrWayback (r=1)	r=15

|> Collaborations and mentorship ● Library, Archive, or Web Archive environment (r=8) ○ Mentorship by library staff (r=3) ○ r: "brainstorming with team members" ○ r: "learning from colleagues" ○ r: "virtual meetings to discuss specific topics between all the personnel dedicated to the [web archive]" ○ r: "Working with colleagues who have a detailed knowledge of web archiving" ○ r: "Working with researchers using archived web data" ● Scholar, Academic, Lecturer, Post-grad/PhD, or working IT/Web Design environment (r=6) ○ r: "collaboration with web archives" ○ r: "Conversations with web archiving service providers and customers" ○ r: "learn a bit from [staff at archive]" ○ r: "Mentor/colleague" ○ r: "Working [...] alongside colleagues in research networks" ○ r: "Working with the team at the [...] Library"	r=14
|> Introductions, guides, manuals ● 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" ● r: "Manual on Gephi" ● r: "Repositories help pages and FAQs" ● Penn Library, Lib Guide: Web Archiving for the Arts and Historic Preservation.	r=4
|> Literature ● r: "books (obviously)" ● r: "Articles by Niels Brügger " ● r: "articles about the history of net art and preserving net"	r=3
|> Information sciences (information studies) ● 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 ● r: "Scaffolding technical skill learning"	r=1
|> Self-learning ● r: "Generally looking up YouTube videos on advanced Excel, Power BI, Gephi etc"	r=1

4.4.5.2 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.

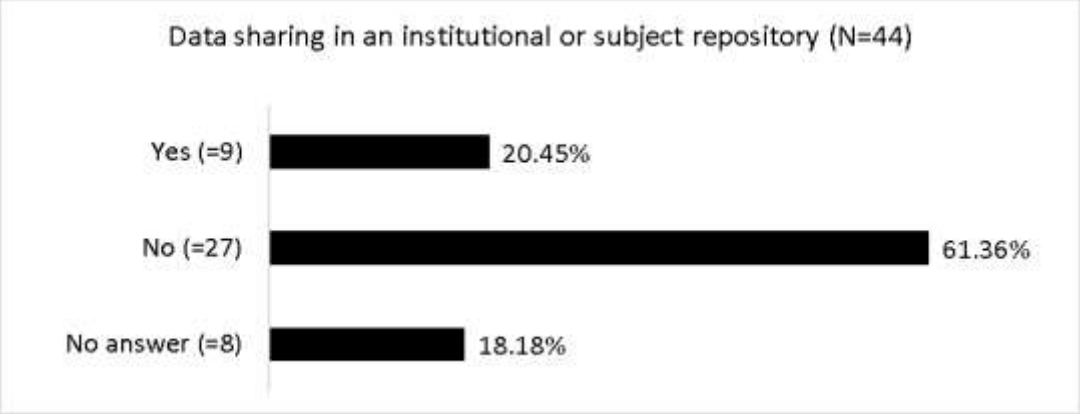

Figure 4.8: Representation of participant responses for whether they shared data in an institutional or library repository (N=44)

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.

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 ● Zenodo, https://zenodo.org (r=2) ● Institut national de l'audiovisuel, https://www.ina.fr/ (r=1) ● Dados.gov +, https://dados.gov.pt/ (r=1)	r=4
|> University repository / library	r=3
|> In-vivo representations ● 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" ● r: "Most data I have shared is via web pages on institutional websites, rather than in specific institutional repositories"	r=2

4.5.6 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 survey 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).

4.5 Discussion

The survey 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, from the survey, 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 survey at least. To add, the participants (N=44) identify with being at novice, intermediate and experienced levels for working with/using web archives.

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

- 4.5.1 - Participants - Positions, Backgrounds, and Interests
- 4.5.2 - Pathways to Web Archive Research
- 4.5.3 - Skills and Knowledge Ecologies in Web Archive Research
- 4.5.4 - Challenges with Web Archive Research

- 4.5.5 - Referencing the Archived Web and Data Sharing
- 4.5.6 - Software, Tools, and Methods used in Web Archive Research
- 4.5.7 - Challenges with Legal Deposit, Copyright, and GDPR
- 4.5.8 - Final Thoughts

4.5.1 Participants - Positions, Backgrounds, and Interests

Regarding the positional background of the participants, we offered two thematic representations: (i) participants who identified with working in a library, archive, or web archive environment (n=30); (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. 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, and 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.

In all, the participants' general interests were varied and diverse across multiple professional fields, practices, specialisms, and academic disciplines as outlined in Figure 4.9. 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 survey identify with one or more of the following subject areas, in alphabetical order (see Figure 4.10).

- 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

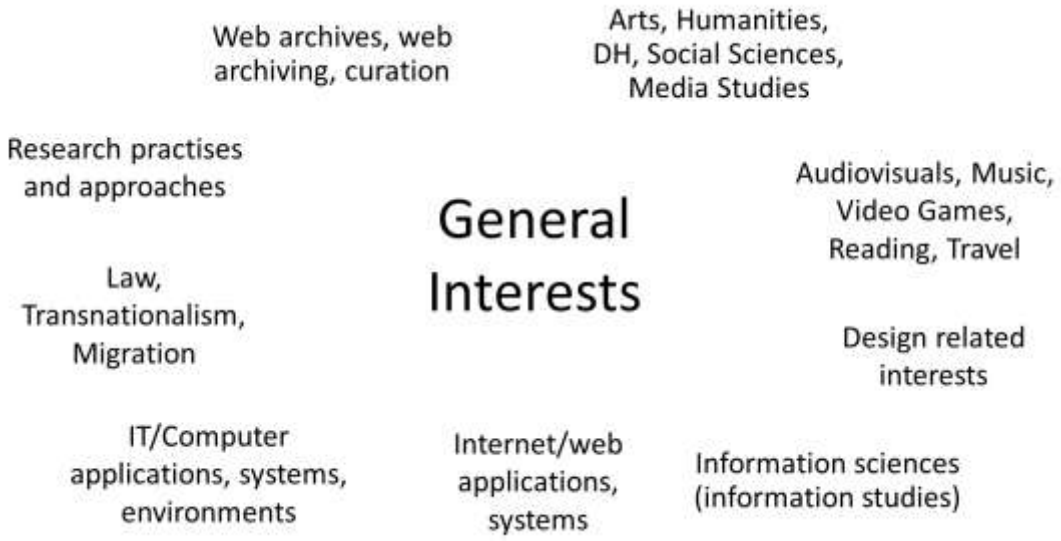

Figure 4.9: WARST participants' interests in general

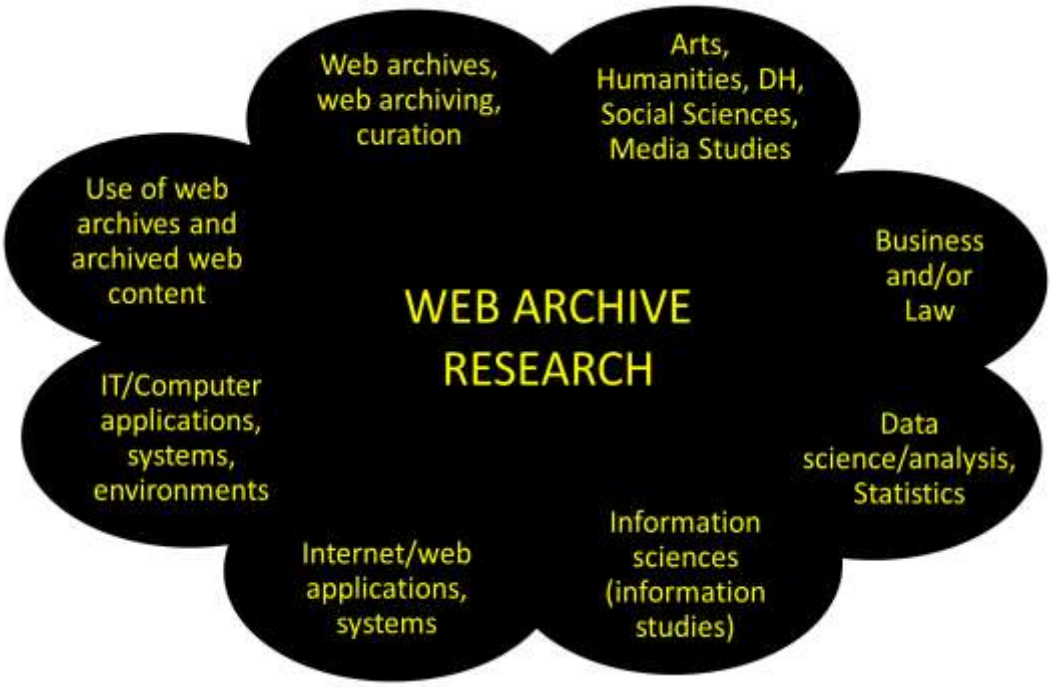

Figure 4.10: In relation to web archive research, the WARST participants identify with one or more of these subject areas

The diversity of participants' background and interests bring to fore the importance of having an interdisciplinary project team conducting this research. The project team included researchers with backgrounds in humanities, digital humanities, cultural studies, media studies, cultural heritage, library and information science, archival science, computer science, and IT development, with different skill sets, areas of expertise, and experiences in working with web archives. This was hugely beneficial for contextualising the participants' responses.

4.5.2 Pathways to Web Archive Research

To better understand the pathways which led the participants to curating/using web archives, we organised two sets of thematic representations from the Results and Analysis and provided 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 4.27, bringing the data together as a whole, with no specific order, or matter of importance. Table 4.27 offers an overview of the thematic representations for the reasons or pathways which led to 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 attempted 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 4.27: 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 ● Concerns about the loss/changes of web content ● Resource to find information/literature ● Business need for a law firm library ● r: "Availability during pandemic" ● Interests in research aspects/outputs of collections ● Digital collection/curation ● r: "It is the present and future of archival work." ● r: "An adviser taught me how to use it." ● r: "My PhD Thesis" ● r: "Internet Archive's Wayback Machine was an early fascination of mine." ● r: "The later development of archival tools to capture and catalog websites has been invaluable" ● r: "A specific collection for a current [...] senator requires capturing his current website" ● Library internship ● Subject librarianship	● Resource for conducting research ● Concerns about the loss of web content ● Resource to find information/old websites ● Business need for web content strategy ● Ease of access to public web archives ● r: "The power of 'raw' internet data to triangulate other data and therefore add to the overall 'scientific' objectivity and credibility of the research" ● Richness of data ● r: "Web archiving is [a] very important topic, which is not researched enough" ● r: "authoritative source" ● r: "Fascination with the centrality of the web in everyday lives and yet its propensity to obsolescence and research oversight" ● r: "Wanting [to] make data available"

4.5.3 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 organised four sets of thematic representations from the Results and Analysis and provided 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 4.28, 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 survey 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 4.28 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 4.28, along with section 4.4.3.5 on the 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, a 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 (2017) 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 4.28: 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 of working with WARCs at scale, and knew almost nothing starting out."			r=1	
r: "Most of my digital skills!"			r=1	

4.5.4 Challenges with Web Archive Research

4.5.4.1 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 pulled together two sets of thematic representations from the Results and Analysis and provided 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 4.29, bringing the data together as a whole, but with no specific order or matter of importance. We further attempted 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 4.29 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 PDFs 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).

Table 4.29: Combined thematic representation of participant responses for challenges encountered in web archive research

Library, Archive, or Web Archive environment (n=25)	Scholar, Academic, Lecturer, Student, or an IT/Web Design environment (n=9)
● Inconsistencies and incompleteness (r=11) ● Legalities for acquisition/access (r=8) ● Challenges with learning new skills (r=6) ● Producing documentation/metadata (r=2) ● Volume of data (r=2) ● Institutional challenges (r=1) ● Technical challenges (r=8) ● Financial challenges (r=4) ● r: "Having access to the raw data, as a web archivist, is very beneficial"	● Inconsistencies and incompleteness (r=10) ● Legalities on access, use, and storage (r=8) ● Challenges with learning new skills (=7) ● Lack of documentation/metadata (r=2) ● Volume of data for research (r=2) ● Challenges in an IT/Business/Admin. environment (r=2) ● Performance related issues (r=1) ● Research methods and approaches (r=5) ● r: "One of the big barriers was getting started" ● r: "once I wanted to get more involved, who to contact!" ● r: "Too many to count!"

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 (2022b) 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) and Costea (2018).

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. 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. This is because it is time-consuming and labour intensive to provide granular metadata. It is also dependent on the availability of financial resources to do so (Costa, 2021, p. 72; Maemura et al., 2018, p. 1226; Jackson, 2015b; 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 that 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 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. 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 constraint, there are also indications from the survey that there would be some benefit in providing users and potential users with introductory web archiving training, localised for the web archive being used. This could raise 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 remarks that the “learning curve was steep”. Another respondent refers to having “Limited technical skills to analyse the WARC-files and the information within them”; another respondent suggests a challenge in “Learning how to use research tools (from a non-

technical user's perspective).” Additionally, 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.

The challenges mentioned above present 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 are 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).

The findings also offer indications 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 signals that staff in a web archiving environment would 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 findings show that participants from a scholarly or academic environment engage with a diversity of tools and methods. However, 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, from the findings 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 / APIs etc.".

Indeed, the findings show several instances which indicate 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 that 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. This would in turn develop their services, thereby benefiting 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 to 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, WAC 2022 programme).[11] However, it is worth noting 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 (2021c) 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) (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 public administrators, journalists, legal

[11] IIPC, WAC 2022 programme, https://netpreserve.org/ga2022/wac/

professionals, web designers, computer scientists, data analysts, and local historians (section 4.4 & 4.5; Ramesh & Hern, 2013; Winters, 2017; Truman, 2016; Bailey, 2015).

4.5.4.2 Comparison between novice, intermediate and experienced levels

To better understand the challenges for web archiving and curation and the use of web archives for 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 4.30 offers a breakdown of 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 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 4.31 offers an overview of thematic representations of 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.

Table 4.30 and Table 4.31 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 diminish with increasing experience highlights the need for training across all levels of capability. 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 4.30: 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 4.31: 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

4.5.5 Referencing the Archived Web and Data Sharing

4.5.5.1 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 note 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)”.

4.5.5.2 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. 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 practices (Zierau et al. 2016; Zierau, 2019).

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 (2019a; 2019b) also describes challenges whereby publicly accessible web archives may be susceptible to link rot if web archive systems change. For example, a web archiving programme may change its service provider or subscription service, as was the case when the National Library of Ireland Web Archive (NLI Web Archive) moved its public selective collections from the Internet Memory Foundation to the Archive-IT platform in 2018 (see Chapter 5). Respondents also identified challenges for citing materials from a legal deposit archive, or an archive with restrictive access, which is problematic for the transparency of the research methods being used. This is further discussed in section 4.5.7. 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.

4.5.5.3 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 indicated a lack of guidelines/standards for citing datasets, and some participants indicated that it is not easy to cite datasets in general. Questions were raised such as (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. Preservation reliability was 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.

4.5.5.4 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 chapter. Truter (2021) offers a comprehensive study focused on this area. As part of the survey, 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 +.

4.5.6 Software, Tools, and Methods Used in Web Archive Research

4.5.6.1 Data collection

To better understand the software and tool ecologies in web archive research, we organised 2 sets of thematic representations for tools and methods used for data collection from the Results and Analysis and provided 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 4.32 presents a comparison of thematic representations for the types of tools and methods used by participants for data collection, and Table 4.33 presents a more detailed breakdown of those tools and methods.

The tables (4.32 & 4.33) 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 4.32: 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) ● Curating web archive collections: selection, configuring and scheduling crawls, annotating seeds, performing QA (r=10) ● Accessing/replaying archived web data (r=8) ● Web archiving subscription services (r=1) ● Collecting data from API (r=2) ● Managing data (r=5) ● Finding source material (r=4) ● Screenshot, screen capture (r=2) ● Web archiving subscription services (r=1) ● Tools with diverse purposes (r=4) (Browser tools, command-line tools, Python scripts/libraries, standard PC tools) ● Digital forensics/preservation (r=1) ● r: "In house developed web archiving tools" ● r: "institutional sources" ● r: "text recognition evaluation tools"	● Crawling software (r=7) ● Curating web archive collections: selection, configuring and scheduling crawls, annotating seeds, performing QA (r=1) ● Accessing/replaying archived web data (r=2) ● Web archiving subscription services (r=1) ● Collecting data from API (r=2) ● Managing data (r=2) ● Finding source material (r=6) ● Screenshot, screen capture, screencast (r=5) ● Web archiving subscription services (r=1) ● Tools with diverse purposes (r=4) (Browser tools, Python scripts/libraries, R (Rstudio)) ● File downloads (r=3) ● Web scraping (extracting data from web pages) (r=2) ● Audio tools (for interviews) (r=1) ● Manual collection for close reading (r=1) ● r: "non-English language search words" ● r: "direct contact with people who might have the data” ● r: "scanning/OCR if the source is hard copy"

Table 4.33: 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) ● ArchiveWeb.page (r=4) ● Browser tools (r=1) ● Browsertrix (r=2) ● Brozzler (r=4) ● command-line tools (r=1) ● Conifer (prior, Webrecorder) (r=9) ● CWeb (r=2) ● DSpace (r=1) ● Electrolyte (r=3) ● Excel, spreadsheet, .csv (=3) ● Heritrix (r=11) ● HTTrack Website Copier (r=1) ● Google Drive (r=1) ● Instaloader (r=1) ● Internet Archive, Wayback machine (r=3) ● Internet, search engines, web search (r=2) ● Library catalogues and databases (r=2) ● MediaArea tools (r=1) ● NetarchiveSuite (r=5) ● OpenWayback (r=2) ● Python scripts/libraries (r=1) ● pywb (r=2) ● screen capture tools (in general) (r=1) ● snipping tools (in general) (r=1) ● Social Feed Manager (r=1) ● Umbra (r=1) ● W3ACT (r=1) ● waybackpy (r=1) ● Web crawler (in general) (r=1) ● Web Curator Tool (r=1) ● Wget (r=1) ● r: "selecting material for collection" ● r: "In house developed web archiving tools" ● r: "institutional sources" ● r: " text recognition evaluation tools" ● r: "the type of tools that come for standard with a PC"	● Archive-It (r=1) ● Audio tools (for interviews etc.) ● Browser tools (r=2) ● Browsertrix (r=1) ● Conifer (prior, Webrecorder) (r=2) ● Heritrix (r=2) ● HTTrack Website Copier (r=1) ● Internet Archive, Wayback machine (r=2) ● Internet, search engines, web search (r=3) ● Library catalogues and databases (r=1) ● Manual collection for close reading ● Manual/scripted file downloads (r=3) ● Python scripts/libraries (r=1) ● R (Rstudio) (r=1) ● screenshot tools/functions (in general) (r=2) ● script for screenshot automation (r=1) ● SHINE tools - UKWA (r=2) ● Snagit (r=1) ● Twarc (=1) ● Web Archiving Integration Layer (WAIL) (r=1) ● Webscraper.io (=1) ● web scraping scripts (=1) ● Websnapper (r=1) ● Wget (r=1) ● Zotero (r=1) ● Zotfile PlugIn (r=1) ● r: "make my own tools to collect data based on [publicly] available API" ● r: "non-English language search words" ● r: "direct contact with people who might have the data" ● r: "scanning/OCR if the source is hard copy"

4.5.6.2 Data analysis

To further examine data analysis, we organised two sets of thematic representations for tools and methods used for data analysis from the Results and Analysis and provided 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 4.34 presents a comparison of thematic representations for the types of tools and methods used by participants for data collection, and Table 4.35 presents 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, standalone 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 citing of tools from the digital humanities and social sciences (Gephi, Voyant Tools, IramuteQ) by some respondents from a library and archive environment points to an ongoing exchange between these communities.

Table 4.34: 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) ● Computer-assisted text analysis (r=2) ● Computing infrastructure (r=1) ● Data extraction, cleaning, transformation (r=6) ● Data management (r=2) ● Digital forensics/preservation (r=3) ● Distributed processing (r=3) ● Evidence analysis (r=1) ● Machine learning (r=1) ● Metadata, crawl logs (r=3) ● Network analysis (r=3) ● Programming/scripting languages, computing environments (r=6) ● Replay/playback tools (r=2) ● Search and information retrieval (r=13) ● Statistics (in general) (r=1) ● Visualisation (r=4) ● Web archive access and analysis (r=1) ● Web archiving management (r=1) ● r: "lists, notes, tiny pieces of paper" ● r: "manual statistics on the report files" from SolrWayback ● r: "My work with the web archive involves selecting material, not carrying out research"	● Collaboration (r=1) ● Computer-assisted text analysis (r=1) ● Qualitative data analysis (r=6) ● Data analysis, extraction, cleaning, transformation (r=8) ● Programming, scripting languages and computing environments (r=8) ● Network analysis (r=3) ● Other Tools (r=3) ● Visualisation (r=1) ● r: "mostly my brain" ● r: "Conceptual tools (e.g. social semiotics, multimodality) for the [analysis] of complex web objects"

Table 4.35: 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) • Amazon Web Services (r=1) • Apache Hadoop (r=2) • Apache Lucene (r=1) • Apache Parquet (r=1) • Apache Solr (r=1) • Apache Spark (r=1) • Archives Unleashed Toolkit (r=1) • BitCurator (r=1) • Collaboration • CDX queries/files (r=2) • Command-line tools (r=1) • Crawl logs (r=2) • Digiboard (r=1) • DROID (r=1) • ElasticSearch (r=1) • Evidence analysis (=1) • Excel, spreadsheets (r=6) • Gephi (r=3) • GLAM workbench notebooks (r=1) • HeidiSQL/MariaDB (r=1) • IramuteQ (r=1) • Jupyter Notebooks (r=1) • Kibana (r=2) • Lucene (r=1) • MediaArea tools (r=1) • Metadata (r=1) • NutchWax (r=1) • OpenWayback (=1) • Python/Python libraries (r=3) • Pywb (r=1) • R (r=1) • SolrWayback (r=2) • SQL (r=2) • Statistics (in general) (r=1) • statistics on the report files from SolrWayback (r=1) • Tableau (r=2) • TensorFlow (r=1) • Voyant tools (r=1) • r: "Web Archive user interface, faceted functions" • r: "lists, notes, tiny pieces of paper" • r: "manual statistics on the report files" from SolrWayback	• Archives Unleashed Cloud (r=1) • Archives Unleashed Toolkit (r=1) • Atlas.ti (r=1) • Bash/shell scripting languages (r=3) • Command-line tools (r=1) • Confluence (r=1) • Excel, spreadsheets (r=4) • Gephi (r=3) • Microsoft 365 (r=1) • Nvivo (r=2) • OpenRefine (r=1) • Pattern matching (r=1) • Proprietary tools (r=1) • Perl (r=1) • Python/Python libraries (r=2) • R (r=1) • Regular expressions (r=1) • Voyant tools (r=1) • r: "mostly my brain" • r: "Conceptual tools (e.g. social semiotics, multimodality) for the [analysis] of complex web objects" • r: "annotating PDFs with PDFExpert" • r: "Close reading of websites and it's html code" • r: "manual qualitative content analysis" • r: "I usually make my own tools" • r: "visualisation tools for qualitative data"

• r: "My work with the web archive involves selecting material, not carrying out research." • r: "brainstorming with colleagues"	

4.5.6.3 Other skills, tools, and methods

Other sections of the findings also present 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, it is 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 a useful skill from novice to advanced levels for the web archive research community.

4.5.7 Challenges with Legal Deposit, Copyright, and GDPR

Across sections of the findings, 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.

Many of the 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 noted 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. 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 (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 practices within legal deposit frameworks, open science frameworks and web archive research environments.

Finally, in section 4.4.3.6 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 4.36. Table 4.36 presents 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 4.36: 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

4.5.8 Final Thoughts

The foregoing shows web archivists need training in all aspects of the web archive lifecycle including an additional set of training on understanding and assessing end user requirements and how to develop training resources to meet these needs (Byrne et al., 2024 forthcoming; Truman, 2016; Gooding et al., 2021). Similarly, individuals who intend using archived web for research or other purposes would require training in multiple areas including how these

collections came into being, and research methods and paradigms that reflect the characteristics of the archived web.

Incidentally, the availability of training on web archiving, or the use of the archived web, is not high on the agenda of courses in higher education which focus on information studies (Byrne et al., 2024 forthcoming). Therefore, it would appear that the value of web archiving and the use of the archived web across academic disciplines as well as policy makers has not yet been fully realised. Byrne et al. (2024 forthcoming) discuss how the development of a collaborative transdisciplinary network would

> be beneficial for the development of educational materials and courses for web archive provisions, but it would also enable discussions to develop frameworks to provide course modules for students in the use of web archives for research, and training courses for educators on how to incorporate web archived content as part of their teaching materials and methods.

There are also parallels between archivists/librarians working with web archive content and the challenges for archivists working with moving image content as pointed out by Lukow (2000) in the report on *Education, Training and Careers in Moving Image Preservation*. Lukow (2000) suggests that

> Aside from the occasional summer school for new archivists, organized by FIAF and held at the Staatliches Filmarchivder DDR, and the equally infrequent single-session university classroom surveys of film preservation, there had been no attempt to offer courses of study in film preservation and film archives administration. The obvious reasons were that there are insufficient positions available for would-be archivists, and most of the work involved in the running of a film [archive] is far removed from the world of academia. The major figures in the American film archival field trained themselves...[they] started at the bottom and slowly graduated to administration (p. 2).

Byrne et al. proffer that in comparison to other areas of information science/management, the web archiving community is still relatively small, and thus, web archiving jobs are few "with some roles including web archiving as part of a wider job profile" (2024 forthcoming). Therefore, they argue "that the universities are being reactive to the job market, rather than proactive to the needs of the academic field, especially when it comes to the issues regarding link rot and the research integrity of URL citations" (Byrne et al., 2024 forthcoming).

To finish here, there has been no widespread movement across academia to address the issue of reference rot, or to advocate for the use of web archiving as a potential. Furthermore, the referencing style guides of organisations such as the American

Psychological Association (APA), Modern Language Association (MLA), and the Institute of Electrical and Electronics Engineers have yet to offer agreed solutions for safeguarding URL citations, even though web archiving provides a somewhat obvious solution. If reference rot were to be taken more seriously by the academic community, researchers too will need to become DIY web archivists to ensure that their research and their bibliographies remain accessible, and thus, more "scientific" (Spinellis, 2003; Byrne et al., 2024 forthcoming).

4.6 Summary

This chapter focused on individuals around the globe who participate in web archive research. It described 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 methodology entailed desk research, participation in WARCnet meeting discussions, and an online survey. The chapter identified and documented the skills, tools, and knowledge required to achieve a broad range of goals within the web archiving lifecycle and explored the challenges for participation in web archive research, and the overlaps and intersections of such challenges across communities of practice (RQ2). It further offered some suggestions and approaches which might be useful for improving the conditions for conducting web archive research (RQ5).

The survey participants (N=44) were aged between 18-64 years, and identified with residing in North America, Europe, and Asia. They acknowledged being at novice, intermediate and experienced levels for working with or using web archives, and there was an equal representation of participants who identified with being male and female. 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 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. For instance, some respondents in the web archiving community indicated that they were also users of various other web archives as part of their workflows and research. Alternatively, some respondents from the scholarly community indicated that they were also creators and curators of web archives for research purposes. Thus, the boundaries between creators, users and technicians are often blurred within 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)

Therefore, this clearly provides some indications of the types of skills, tools and knowledge that are necessary for conducting web archive research.

The findings demonstrated 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 technologies keep evolving, so too will the challenges. They further highlighted how 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 (Winters, 2020a; Hock-Yu, 2014; Gooding et al., 2021; Vlassenroot et al, 2019; Graham, 2019; Brügger, 2021c; Ogden, 2021; Ogden et al. 2022; Ben-David, 2021). Therefore, it is imperative that creators and users/researchers keep moving forward as collaborators to guide the next generation of web archive research. How this relates to web archive research in Ireland, or how useful web archives are as resources for conducting Irish based research, has yet to be examined, and will be considered in more detail in the following chapters. As a starting point, it would be worthwhile examining the availability and accessibility of web archives which would prove useful for conducting Irish based research.

5.0 REVIEW OF WEB ARCHIVES FOR IRISH BASED RESEARCH

> *Finding a balance between preservation and access is the most urgent problem to be solved, because if today's Web is not saved it will not exist in the future. Access is a political as well as a legal problem. The answer to the access problem, like the answers to all political problems, lies in establishing a process of negotiation among interested parties. Who are the stakeholders, and what are the stakes, in building a Web archive? (Lyman, 2002, p. 40).*

5.1 Introduction

The previous chapter discussed the challenges for participation in web archive research and how these challenges overlap and intersect across communities of practice. It concluded 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. It further demonstrated the need for ongoing collaborations between web archive creators and users to guide the next generation of web archive research. This is a useful starting point when it comes to examining the current state of web archive research in Ireland, or how useful web archives are as resources for conducting Irish based research, which has been relatively understudied.

This chapter forms part of a collaborative investigation by Sharon Healy (Maynooth University) and Helena Byrne (British Library) and uses a qualitative exploratory approach, through desk research, a review of the literature, and informal dialogues with heritage colleagues to offer an overview of the availability and accessibility of web archives based on the island of Ireland, and their usefulness as resources for Irish based research. In doing so, it examines some of the causes for the loss of Irish digital heritage (RQ1), the challenges for participation in web archive research in Ireland (RQ2) and the accessibility and availability of web archives based on the island of Ireland for conducting research on Irish based topics (RQ3). The chapter also offers some perspectives with regard to improving the conditions for conducting web archive research (RQ5). As mentioned previously, we refer to Irish digital heritage in the context of the island of Ireland. When required, we will refer to the digital heritage of Northern Ireland or the Republic of Ireland to distinguish between the two jurisdictions.

At present, there are three main web archiving initiatives which capture websites as part of their efforts for the preservation of digital heritage for the island of Ireland. The Public

Record Office of Northern Ireland (PRONI) takes responsibility for capturing websites related to the six counties of Northern Ireland through a selective collection approach and provides online access to their collections via the PRONI Web Archive.[12] The UK Web Archive also has responsibility for capturing websites in Northern Ireland through their annual national web domain crawl, under The Legal Deposit Libraries (Non-Print Works) Regulations 2013 (NPLD).[13] To note here, as a UK Web Archive partner, the Library of Trinity College Dublin (TCD) provides onsite access to the UK Web Archive's NPLD national web domain collections through their onsite library PC system (Library of TCD, n.d., Electronic Legal Deposit). The National Library of Ireland (NLI) takes responsibility for capturing websites, as part of its strategy for preserving the digital heritage of the twenty-six counties in the Republic of Ireland, through a selective collection approach which is accessible online through the NLI Web Archive.[14] The NLI Web Archive have also conducted three national web domain crawls in 2007, and 2017. However, these collections are currently inaccessible to researchers or the public due to legislative matters and will be discussed in more detail further on.

There is no doubt that Irish digital heritage can be found in other web archives such as the Internet Archive (Wayback Machine), Common Crawl, and Archive.today.[15] Other web archives which have relevance to Northern Ireland are the UK Government Web Archive and the UK Parliament Web Archive. The UK Parliament Web Archive captures, preserves and makes accessible "UK Parliament information" that is published on the web, and the "web archive includes UK Parliament websites and social media dating from 2009 to the present" (UK Parliamentary Archive, n.d., UK Parliament Web Archive).[16] The UK Government Web Archive captures, preserves, and makes accessible "UK central government information published on the web. The Web Archive includes videos, tweets, images and websites dating from 1996 to the present day" (The National Archives, n.d., UK Government Web Archive).[17]

There may also be other minor web archiving initiatives being conducted which have relevance for Irish heritage, by researchers in an academic setting, or for business purposes. For example, the Library of University College Dublin (UCD) collected circa 150 websites "relevant to Irish poetry in the 21st century" which are hosted, and openly accessible on the

[12] PRONI Web Archive, https://webarchive.proni.gov.uk/#!/

[13] UK Web Archive, https://www.webarchive.org.uk/ukwa

[14] NLI Web Archive, https://archive-it.org/home/nli

[15] Wayback Machine (Internet Archive), https://archive.org/web; Common Crawl, https://commoncrawl.org; Archive.today, https://archive.ph

[16] UK Parliament Web Archive, http://webarchive.parliament.uk/atoz

[17] UK Government Web Archive, https://www.nationalarchives.gov.uk/webarchive/

Archive-It service platform.[18] As the collection strategy for the library's Irish Poetry Reading Archive has become increasingly digital, with "a huge amount of material of importance" only being available on the web, it embarked on a project to collect and preserve websites relevant to Irish poetry for future generations (UCD Collections, Archive-It). Other examples include social media archive collections.

Such archives include a collection by Darcy et al. (2021) in the Digital Repository of Ireland, titled 'In Her Shoes: Stories of the Eighth Amendment', which was collected as one part of a wider collection programme by the award winning Archiving Reproductive Health project (Digital Repository of Ireland, 2022).[19] The collection comprises administrative posts and 'stories' which were published on the Facebook page, 'In Her Shoes - Women Of The Eighth', during the run up to the Referendum to the Repeal the Eighth Amendment, colloquially known as the 'abortion' referendum which took place in the Republic of Ireland on 25 May 2018.[20] The referendum was successful, resulting in a constitutional change through the Thirty-sixth Amendment of the Constitution Act 2018.[21] There are also other archived social media datasets which relate to the Repeal the Eighth referendum. These include Littman's (2018) 'Ireland 8th Tweet Ids', which was collected from the Twitter filter stream API using Social Feed Manager and is available in the Harvard Dataverse repository[22], and Ó Briain and Foster's (2020) '#retweetthe8th: twitter dataset' which was collected from the Twitter filter stream API using Twarc, between 09 March and 30 May 2018, and is available in the Zenodo repository.[23]

For this chapter, however, we are interested in web archive initiatives that are based on the island of Ireland which have a specific mandate to capture a wide range of Irish digital heritage as part of their collection development strategies. Therefore, we focus on the

[18] UCD Special Collections, Archive-It, https://archive-it.org/organizations/1846

[19] DRI, In Her Shoes: Stories of the Eighth Amendment, https://repository.dri.ie/catalog/wm11nd02p

[20] Eighth Amendment of the Constitution Act, 1983, https://www.irishstatutebook.ie/eli/1983/ca/8/enacted/en/html?q=Eighth+Amendment+of+the

[21] Thirty-sixth Amendment of the Constitution Act 2018, https://www.irishstatutebook.ie/eli/2018/ca/36/enacted/en/html

[22] Ireland 8th Tweet Ids, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PYCLPE

[23] Ireland 8th Tweet Ids, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PYCLPE; #retweetthe8th: twitter dataset from the 2018 Referendum to repeal the 8th Amendment of the Constitution of Ireland, https://zenodo.org/record/3842013#.YZOx_S2l1bV

PRONI Web Archive, the NLI Web Archive, and the UK Web Archive, which is accessible onsite in the Library of TCD; and we examine the availability and accessibility of these resources for conducting research on Irish based topics. In doing so, we offer an overview of these web archiving initiatives and their historical backgrounds, inclusive of how copyright and legal deposit has influenced their collecting activities. We further observe their efforts for the collection and preservation of digital heritage from the web spaces of Northern Ireland (NI) and the Republic of Ireland (ROI), and we assess their availability and accessibility as resources for conducting Irish based research. As mentioned previously, this chapter refers to Irish digital heritage in the context of the island of Ireland, but it is also worthwhile examining some of the historical context in which the heritage of Ireland was preserved preceding the digital turn.

5.2 Preservation of Irish Records and Publications, pre-Digital

Prior to the Irish War of Independence (1919-21), the Anglo-Irish Treaty (1921) and the 1920-21 partition of the island of Ireland, under UK legislation the NLI was a nominated institution for the collection of books and publications, while the State Paper Office and the Public Record Office of Ireland were the nominated institutions for the preservation of the state records of Ireland. Following partition, PRONI was established "for the reception and preservation of public records appertaining to Northern Ireland" (Section 1: Public Records Act (Northern Ireland), 1923). In the next section, we briefly look at some of the backgrounds of these institutions.

The NLI was established in 1877, following negotiations between the Royal Dublin Society, the Department of Science and Art (London) and the Commissioners of Public Works (Ireland), which led to the Dublin Science and Art Museum Act of 1877 and the establishment of a national library and national museum. As Ireland was part of the UK at the time, the newly established NLI was governed from London up until the establishment of the Irish independent state in 1922, when it was handed over to the Irish Government under the remit of the Department of Education. From July 1986, the NLI was transferred to the Department of An Taoiseach, and transferred again in 1992 to the newly established Department of Arts, Culture and the Gaeltacht (NLI, n.d., History of the Library).

The National Cultural Institution Act, 1997 confirmed the NLI as the official library of record in the ROI in its responsibility for collecting for, and on behalf of the Irish state (NLI,

Collection Development Policy 2022-2026, p. 2).[24] In 2005, the NLI became an autonomous agency, governed by a Board, and is currently under the aegis of the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media (previously called the Department of Culture, Heritage and the Gaeltacht) (NLI, n.d., History of the Library; Collins, 2018).

Today, the mission of the NLI is to "collect, protect and make accessible the recorded memory of Ireland" inclusive of "books, serials, newspapers, manuscripts, maps, photographs, official publications, prints, drawings, ephemera, digitised and born-digital collections" (NLI, Collection Development Policy 2022-2026, p. 1). In doing so, it is also committed to collecting "representatively and inclusively" in order "to capture the diversity of Irish experience" (NLI, Collection Development Policy 2022-2026, p. 2) and "create a more diverse and inclusive story of Ireland, so that new voices are collected and shared with the world" and ensure "that Ireland is represented in all its diversity, in all of our activities and that equal access to these is provided for everybody" (NLI, Diversity and Inclusion Policy 2018-2021, p. 1). Thus, from 1887 to the current day, the NLI has been a leading force in the preservation of Irish heritage.

In terms of state records, a State Paper Office (SPO) for Ireland was established in 1702, to preserve the records of departing chief governors, and the keeping of "any Kings' or Queens' Letters, warrants, orders, petitions and other letters belonging to the Secretaries' Offices" (Wood, 1930, p. 23). Administered from Dublin Castle, the SPO was handed over to the new independent Irish State in January 1922 (Maguire, 2022) and would continue to operate until the 1980s. The Public Records (Ireland) Act, 1867 provided for the establishment of a Public Record Office of Ireland (PROI) in a purpose-built facility in the Four Courts complex in Dublin, consisting of a "Record House" and a "Record Treasury" (Wood, 1930, pp. 30–33).[25]

According to the National Archives of Ireland website, the buildings

> consisted of a three-storey over-basement Record House with staff offices, a caretaker's apartment, a library, a binding room and a public reading room. Behind the Record House was the Record Treasury, an enormous six-storey building containing 100,000 square feet of shelving with records accumulated over seven centuries (National Archives of Ireland, n.d., Public Record Office of Ireland).

[24] National Cultural Institutions Act, 1997, https://www.irishstatutebook.ie/eli/1997/act/11/enacted/en/html

[25] Public Records (Ireland) Act, 1867, https://www.nationalarchives.ie/PDF/PROI1867.pdf

The transfer of records to the new PROI facility included records from Dublin Castle, the Landed Estates Record Office at the Custom House, and legal records held by the courts in the Four Courts complex (Wood, 1930, pp. 29–31). Moreover, "Every year the documents in the various public offices throughout Ireland which had arrived at twenty years of age were automatically transferred to the Record Office" (Wood, 1930, p. 33). In June 1922, the Record Treasury of the PROI was destroyed by fire and explosion in the opening days of the Irish Civil War (Wood, 1930, p. 35). Regan (2016) posits that the destruction of the records of the PROI marked "a cultural atrocity unique in modern Irish history" (p. 11). By 1928, the PROI was again open to the public, and "herculean efforts were made by the staff to find replacements for records that had been destroyed" (Crowe, 2012). These efforts were further assisted by the establishment of the Irish Manuscript Commission in 1928, who were given the remit of reporting on

> the nature, extent, and importance of existing collections of manuscripts and papers of literary, historical and general interest relating to Ireland, and on the places in which such manuscripts and papers are deposited, and to advise as to the steps which should be taken for the preservation and publication of such manuscripts and papers, whether in public collections or in private ownership (Dáil Éireann, Ceisteanna—Questions. Oral Answers - Manuscripts Commission, 17 October 1928; Irish Manuscripts Commission, n.d., Mission).[26]

The PROI and SPO continued to function until the enactment of the National Archives Act, 1986, which established the National Archives of Ireland on 1 June 1988. The Act transferred both the functions and holdings of the PROI and the SPO to the newly established National Archives.[27] Furthermore, "Under this legislation, records of Government Departments and their agencies are transferred to the National Archives when they are 30 years old" (NAI, n.d., About the National Archives). The thirty-year rule will be gradually decreased to a twenty-year rule in the foreseeable future (McGee, 2018).

In 1989, the NAI were assigned new premises in Bishop Street, Dublin, and the SPO in Dublin Castle was vacated in August 1991, and the PROI facility in the Four Courts was vacated in September 1992, as the NAI began operations from their new headquarters in Bishop Street (NAI, n.d., About the National Archives). The "salved records" which had been saved from the rubble of the PROI Record Treasury were finally dealt with as part of an innovative

[26] Dáil Éireann. (1928). Ceisteanna—Questions. Oral Answers. - Manuscripts Commission, https://www.oireachtas.ie/en/debates/debate/dail/1928-10-17/3

[27] National Archives Act, 1986, https://www.irishstatutebook.ie/eli/1986/act/11/enacted/en/html?q=National+Archives+Act

project titled ‘Beyond 2022: Virtual Record Treasury of Ireland’. According to the project’s website,

> Beyond 2022 is an all-island and international collaborative research project working to create a virtual reconstruction of the Public Record Office of Ireland, which was destroyed in the opening engagement of the Civil War [...] Together with our 5 Core Archival Partners and over 40 other Participating Institutions in Ireland, Britain and the USA, we are working to recover what was lost in that terrible fire one hundred years ago (Beyond 2022, n.d., Home)

Through the identification of duplicate documents in archives elsewhere, or documents which may reference such documents, and the conservation of the “salved records” from the debris, the project was able to digitise, transcribe, and assign metadata to a rich assortment of replacement materials. Coinciding with the centenary of the Four Courts catastrophe at the end of June 2022, the Beyond 2022: Virtual Record Treasury of Ireland was launched online, bringing together the replacement materials within an immersive 3-D (re)construction of the destroyed Record Treasury building (Figure 5.1). The project was funded by the Irish Government under Project Ireland 2040, through the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media (Beyond 2022, n.d., Support).

The Beyond 2022 project is a great example of how to create academic and popular awareness of a heritage problem while using technology to provide solutions. The project’s efforts to identify paper documents from multiple archives and libraries, digitise such documents and assemble such documents as replacement documents for those that were destroyed in the PROI, offers a model which could be used in other countries where archived collections have been destroyed due to fire, flooding or natural disasters for example. However, it should be stressed that it would not be much use as a model for born digital heritage such as websites, as web content often “replaces its antecedent, usually leaving no trace of the previous document/edition”, and once a web document is “permanently removed from the WWW it ceases to exist” (Koehler, 1999). Thus, unless it is archived, it is lost to perpetuity.

While the NAI in its current form was established in the late 1980s, at the dawn of the digital age, it has had significant challenges when it comes to digital heritage, as outlined in section 2.3. From at least 1997, the NAI continually warned the Irish government regarding the loss of Irish digital heritage, through the loss of electronic records due to obsolete technology, or the fact that there is a lack of formal record keeping guidelines for electronic information by government departments and agencies. Section 2.3 also demonstrated how content on the Irish government website(s) has changed and disappeared over the past decades. However, the NAI is not currently involved in a public web archiving initiative for the

websites of the government, its departments, or associated agencies. While the NAI produced an ambitious strategy in 2021 to deal with the information age including "a digital transformation programme [and] a new framework for records management across government" (NAI, n.d., News; NAI, 2021b), section 2 highlighted that the strategy will depend on "improved funding, an enhanced infrastructure and [...] improved staffing resources" (NAI, 2021b, p. 6). As the NAI gradually moves from a thirty-year rule to a twenty-year rule for the transfer of government records to the NAI (McGee, 2018), it is already facing the problem of curating the first wave of digital records from government departments, which is bound to increase exponentially.

Figure 5.1: Screenshot of the interface of the 3D Virtual Record Treasury of Ireland (https://vrtour.virtualtreasury.ie), taken on 2022-10-18[28]

Regarding NI, following partition, the Public Record Office of Northern Ireland (PRONI) was established through the Public Records Act (Northern Ireland), 1923. As its purpose it would serve as an archive "for the reception and preservation of public records appertaining to Northern Ireland which otherwise would be deposited in the Public Record Office of Ireland" (Section 1: Public Records Act (Northern Ireland), 1923). The Act also permitted PRONI to collect documents which did not expressly relate to Government creation or use (PRONI, 2008). PRONI combined the functions of a public and a state records office.

[28] Beyond 2022, Virtual Tour: Record Treasury of Ireland, https://vrtour.virtualtreasury.ie/

PRONI began operations in March 1924 in a disused linen warehouse in Murray Street, Belfast. Under the guidance of the first Deputy Keeper, Dr D.A. Chart, their first mission was to find and recover surrogates of records which had been destroyed by fire and explosion at the Public Record Office of Ireland in Dublin during the Irish Civil War (PRONI, 2007; PRONI, 2008). Having previously worked at the Public Record Office of Ireland in Dublin, Dr Chart was familiar with the records which had been destroyed, and immediately set out to solicit duplicate records held by churches, solicitors, politicians, businesses, and the landed aristocracy – which proved to be very fruitful (PRONI, 2007; PRONI 2008). In April 1933, PRONI moved to the newly built Courts of Justice building, in Chichester Street, Belfast, and moved again in 1968 to a purpose-built location in Balmoral Avenue, Belfast (PRONI, 2007). PRONI moved once more in 2011 to its current location, at a purpose-built facility in the Titanic Quarter of Belfast (NIdirect, n.d., Getting to PRONI).

From 1924, PRONI came under the jurisdiction of the Ministry (later Department) of Finance and moved to the jurisdiction of the Department of the Environment in 1982 where it became an executive agency within that department. From 1999, PRONI became an agency under the Department of Culture, Arts and Leisure with the restoration of a devolved government. From 2006, PRONI ceased to be an agency, and became a division within the main Department of Culture, Arts and Leisure (PRONI, 2008).

Mostly dating from the seventeenth century to the present day, PRONI contains millions of records "that relate chiefly, but by no means exclusively, to Northern Ireland" (CAIN, n.d.). Their holdings fall under the following categories:

- privately deposited archives: e.g., landed estate archives, business records, church registers, emigrant letters, etc.
- public records: e.g., records from official sources such as local authorities, courts of law, quangos, public bodies, etc.
- departmental records of the various departments and ministries involved in the governance of Northern Ireland since 1921 to present (CAIN, n.d.).

5.3 NI Web Space

5.3.1 PRONI Web Archive

In NI, PRONI began a selective web archiving programme circa 2010, to capture and preserve websites of government departments, local councils, public sector organisations and websites "of social, political, cultural, religious or economic significance and relevance to Northern Ireland" (NIdirect, 2015; Murchan, 2020a). The resource is publicly available online

as the PRONI Web Archive. On the technical side, PRONI partners with Archive-It, a subscription-based web archiving service provided by the Internet Archive in the United States (US). However, it originally partnered with the Internet Memory Foundation (IMF), a subscription-based web archiving service in Europe, which ceased operations circa August 2018 (NIdirect, 2015; Wikipedia, 2012+, Internet Memory Foundation; Aturban, 2019b).

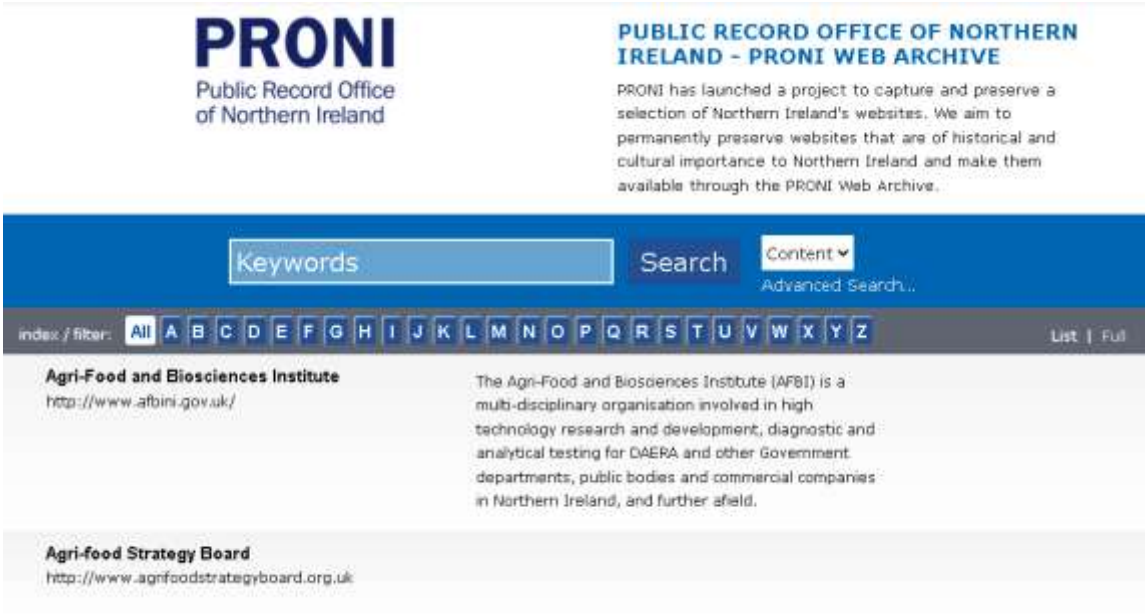

Figure 5.2: Screenshot of the interface of the PRONI Web Archive, taken on 2022-08-24

PRONI operates a two-fold approach to selective collection. First, regarding websites under the jurisdiction of NI government departments, arm's-length bodies, or other publicly funded bodies, PRONI "does not seek permission to crawl and publish" such websites, rather it notifies the website owners/government departments of their intentions to include the website in their collection and provides a takedown policy (Murchan, 2020b; PRONI, 2018, p. 5). The takedown policy also accounts for cases whereby a website, or parts of a website, may be in dispute due to third party copyright (e.g., copyrighted photographs), or in breach of data protection due to the availability of an individual's personal information, for example (PRONI, 2018, p. 5; PRONI, 2016). With the second approach, PRONI does, however, seek permissions "to crawl and publish privately funded" websites which are deemed historically or culturally important for inclusion in the PRONI Web Archive (PRONI, 2018, p. 5). In August 2022 there were 320 websites listed as available, of which many have accompanying descriptive metadata on the interface (see Figure 5.2 & Figure 5.3). The resource can be searched by inputting a known URL, and through a search of the content with keywords. There is also an advanced search option which includes full-text or phrase search and refining the search through date parameters (Figure 5.4).

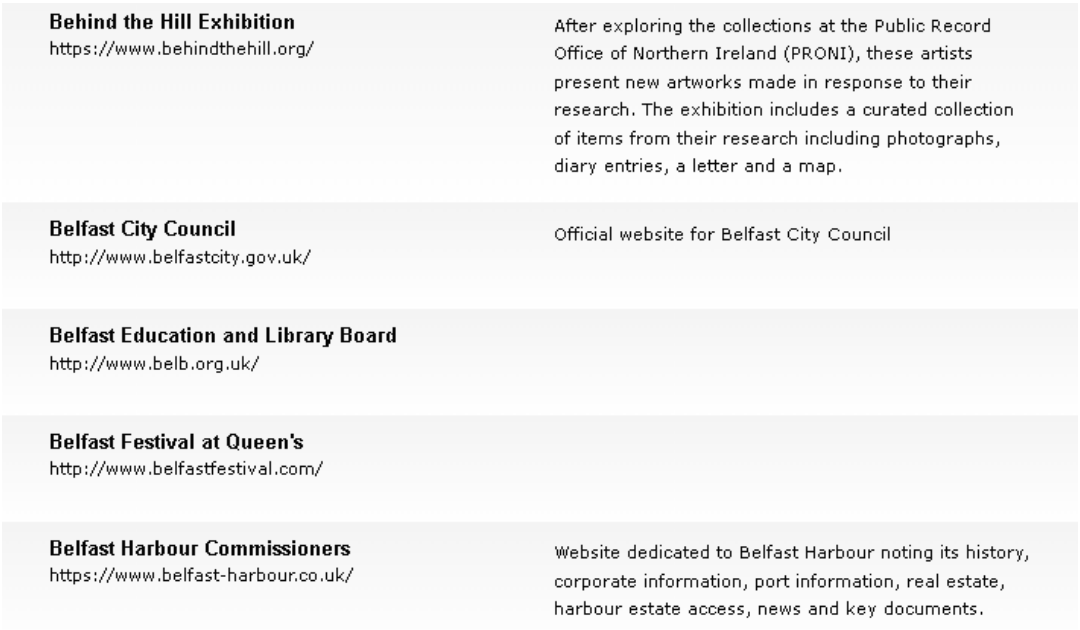

Figure 5.3: Screenshot of the interface of the PRONI Web Archive, showing some descriptive metadata entries, taken on 2022-09-26

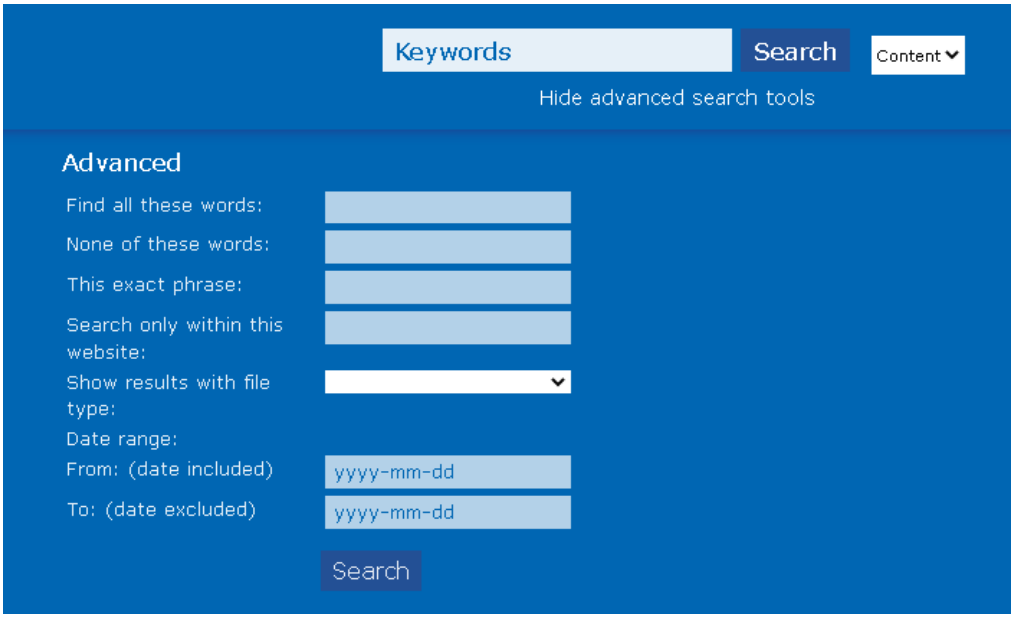

Figure 5.4: Screenshot of the advanced search options interface of the PRONI Web Archive, taken on 2022-10-02

5.3.2 UK Web Archive

In addition to this, UK NPLD allows for the archiving of the NI web space under the jurisdiction of the UK legal deposit libraries. Access to the NPLD web archive domain collection is only available onsite in a legal deposit library. These include the British Library,

the National Library of Scotland, the National Library of Wales, the Bodleian Library at Oxford University, Cambridge University Library, and the Library of Trinity College Dublin (The Bodleian Libraries, n.d., Legal Deposit).

Legal deposit for print publications first became a part of English law through the Licensing of the Press Act (1662) and became a more formalised British law through the Copyright Act, 1710, also known as the Statute of Anne (Partridge, 1938, p. 33; Koehler, 2015, pp. 149–150). The Copyright Act, 1710 formalised a relationship between copyright and legal deposit as part of the same legislation. However, legal deposit was only officially extended to Ireland with the enactment of the Act of Union of Great Britain and Ireland in January 1801.[29] This also formalised the relationship of the Library of TCD as a UK legal deposit library up to the current day (Library of TCD, n.d., Electronic Legal Deposit; Partridge, 1938, p. 45). Partridge (1938) suggests that it was a great failure on the part of the Westminster Parliament not to include Ireland in the Copyright Act 1710, as Ireland became a haven for piracy with the reprinting of English and Scottish books throughout the 1800s (p. 134).

During the 1800s there were several more changes to legal deposit with the Copyright Acts of 1814 and 1836 which first increased and then decreased the number of nominated legal deposit libraries (Muir, 2005, p. 14). Thereafter the 'Imperial' Copyright Act of 1842 was introduced as an attempt to regulate copyright legislation throughout the British empire (Partridge, 1938, p. 80). Partridge (1938) notes that the Act of 1842 was designed to ensure that the British Museum (later renamed British Library) secured a copy of "every book anywhere under British rule" (p. 80), and copyright law then remained practically unchanged for much of the century (Feather, 1994, p. 6). The Copyright Acts of 1814, 1836 and 1842 also sought to recognise authors, who had up to this point been somewhat neglected by the Copyright Act, 1710 (Feather, 1994, p. 6).

In the 1900s, the Copyright Act 1911 extended the copyright term of an author and incorporated the National Library of Wales as a legal deposit library (Partridge, 1938, p. 108), although it could only claim "material in Welsh or of Welsh or Celtic interest" which would not change until the 1970s (Feather, 1994, pp. 120–121). The 1911 Act stipulated that a copy of print books be deposited, free of charge, in the British Museum within a month of publication, and the other five legal deposit libraries could claim a copy within 12 months of publication (Muir, 2005, p. 19). The five legal deposit libraries included the National Libraries of Scotland and Wales, and the University libraries of Oxford, Cambridge, and Trinity College Dublin (Working Party on Legal Deposit, 1998). Thereafter, the Copyright (British Museum)

[29] Act of Union (Ireland) 1800, https://www.legislation.gov.uk/aip/Geo3/40/38/contents

Act 1915, reduced the types of publications to be deposited in the British Museum, exempting such items as rail timetables, advertisements, voter rolls, design specifications, and calendars (Muir, 2005, p. 18).

Hence, from 1801 up until the establishment of the independent Irish State in 1922, Ireland came under the jurisdiction of UK copyright and legal deposit laws. Thereafter, they only applied to the six counties of NI. It would then take until 1927 for the Irish State to introduce its own legislation on copyright and legal deposit through the Industrial and Commercial Property (Protection) Act, 1927.[30]

There were further changes to UK copyright and legal deposit laws such as The Theatres Act 1968 which allowed for the incorporation of published theatre production scripts as part of legal deposit, although there was a failed attempt to extend legal deposit to films through the Film (Statutory Deposit) Bill in 1969 (Muir, 2005, p. 18).[31] Also, copyright was amended in the Copyright Act, 1956, but it did not affect legal deposit (Muir, 2005, p. 20). Muir (2005) suggests that for legal deposit the Copyright Act, 1911 remained relatively unchanged until the end of the century (p. 23).[32] Nonetheless the inclusion of non-print material cropped up on the agenda from time to time. Muir (2005) notes how several proposals were put forward to extend legal deposit to non-print material such as "microfilm, sound and audiovisual material, such as films" (p. 20). In addition, as the UK was a member of the European Economic Community (EEC) since January 1973, which was later renamed the European Community (EC), and thereafter the European Union (EU), the UK was often subject to directives which attempted to harmonise copyright across member states (The National Archives, n.d., The EEC). For example, EEC member states had to comply with a 1993 Council Directive, which was an attempt to harmonise copyright across member states on account of the

> differences between the national laws governing the terms of protection of copyright and related rights, which are liable to impede the free movement of goods and freedom to provide services, and to distort competition in the common market (Council Directive No. 93/98/EEC (2)).

In terms of legal deposit, however, it was not until the mid-1990s that the legislation would really come under scrutiny, due to the "development of new media and the growth of publication in non-print forms" (Working Party on Legal Deposit, 1998). The growth of born

[30] Industrial and Commercial Property (Protection) Act, 1927, https://www.bailii.org/ie/legis/num_act/1927/0016.html#zza16y1927

[31] Theatres Act 1968, https://www.legislation.gov.uk/ukpga/1968/54

[32] Copyright Act 1911, https://www.legislation.gov.uk/ukpga/Geo5/1-2/46/introduction/enacted

digital media could not be denied, due to the rapid development of the web, and the increasing availability of the internet (Masanès, 2006, p. 3).

A proposal to the UK Government to extend legal deposit to non-print materials including digital material was submitted by the British Library in 1996, on behalf of the legal deposit libraries and the British Film Institute. This kick-started a commitment by the UK Government to develop a legal deposit scheme for non-print, with the setup of an interim voluntary scheme for microfilm in 2000, pending forthcoming legislation (Muir, 2005, p. 4; p. 29). In taking the UK into the digital information age, the Legal Deposit Libraries Act 2003 extended the UK legal deposit scheme to "non-print (electronic) publications, including websites, subject to further enabling Regulations" (UK Web Archive, n.d., FAQ).[33] These Regulations would not come into effect until 2013 through The Legal Deposit Libraries (Non-Print Works) Regulations 2013 (NPLD).[34] The 2003 Act also separated copyright law from legal deposit law (Muir, 2005, p. 3). Nonetheless, the 2003 Act did provide the legislative framework to enable the archiving of web content, albeit with permission from website owners (Bingham & Byrne, 2021, p. 2), and mandated the responsible minister with the powers to bring in regulations for digital collecting, including websites, under legal deposit which could be enacted at the appropriate time in the future.

Following a report commissioned by the Wellcome Trust and the Joint Information Systems Committee (JISC) (Day, 2003) on the feasibility of developing a UK web archiving service, six institutions came together to form UKWAC "to experiment with collection of website materials before the implementation of legal deposit legislation covering web publishing" (UK Web Archive, 2009, F.A.Q.; Bingham & Byrne, 2021, p. 2). The institutions included: The National Archives, the British Library, the national libraries of Scotland and Wales, the Wellcome Library and JISC (Bailey & Thompson, 2006). The UKWAC web archive was publicly launched in May 2005 with some of its earliest collections being the Indian Ocean Tsunami December 2004, the 2005 General Election and the 2005 July London terrorist attacks (Bailey & Thompson, 2006; UKWAC, 2005; UK Web Archive, 2020).[35]

It should also be noted however, that publishing, and communications technologies had rapidly advanced between the time the 2003 Act came into force in early 2004 and the time the 2013 NPLD Regulations came into effect (Arnold-Stratford & Ovenden, 2020, p. 5). From

[33] Legal Deposit Libraries Act 2003, https://www.legislation.gov.uk/ukpga/2003/28/contents

[34] The Legal Deposit Libraries (Non-Print Works) Regulations 2013, https://www.legislation.gov.uk/uksi/2013/777/contents/made

[35] With the passing of the 2013 NPLD Regulations, the UKWAC collections automatically transferred to the UK Web Archive under the partnership of the six UK legal deposit libraries.

2003 – 2013 the practical details for the NPLD regulations were worked out, during which time the UK Legal Deposit Libraries (LDLs) selectively archived websites under existing copyright law while contributing to the discussion on whether digital collecting needed legislation or whether it could be carried out under voluntary deposit. The conclusion was that seeking permission to archive from website publishers was not feasible and the regulations were necessary. The legislation was therefore updated in 2013 to allow for an annual web crawl of the UK web estate, including NI, undertaken by the UK Web Archive, a partnership of the six UK LDLs. The NPLD Regulations also solidified the establishment of the UK Web Archive (Bingham & Byrne, 2021, p. 2).

In terms of collection, the UK Web Archive offers the following summary regarding their legal deposit collection strategy

> As per the Non-Print Legal Deposit regulations we the six UK Legal Deposit Libraries are empowered to collect any and all UK based websites. In effect this includes all websites that have a UK top level domain name such as .UK, .SCOT, .WALES, .CYMRU and .LONDON plus any websites that are identified as being hosted on a server located physically in the UK via a geo-ip lookup. Additionally, if a website contains a UK postal address or the website owner confirms UK residence or place of business their website can be included. In order to build comprehensive thematic website collections, we occasionally request permission to archive non-UK websites from the site owner (UK Web Archive, n.d., Frequently Asked Questions).

Thus, the extent of the scope of their legal deposit collection efforts goes well beyond the capture of a national web domain demarcated by a basic ccTLD. Nonetheless, there are still "significant gaps in the heritage acquired as websites on 'non-UK' top level domain names, such as .com, are not automatically identified" (Bingham & Byrne, 2021, p. 2). This highlights the challenges for demarcating the geographical, structural and "imaginary" boundaries of a national web domain (Ben-David, 2019, pp. 89–91; Kahn, 2019, pp. 164–165; Webster, 2019, pp. 110–112).

In addition, the PRONI Web Archive commenced a selective web archiving initiative in 2010 to capture and preserve websites of NI government departments, local councils, public sector organisations and websites which have social, cultural, political, religious, or economic significance for the preservation of NI heritage. However, prior to 2013, the UK/NI web space was not systematically captured as part of legal deposit, and therefore much of the earlier NI webspace will have disappeared or changed drastically (Jackson, 2015a). To salvage some of the UK web estate prior to 2013, the Joint Information Systems Committee (JISC) acquired a dataset from the Internet Archive which included all .uk websites in their

web archive collections that were crawled from 1996-2013 (UK Web Archive, n.d., JISC UK Web Domain Dataset).

The UK Web Archive is managed by the British Library, inclusive of its technical infrastructure (Pennock, 2013, p. 27). Initially, UKWAC utilised PANDAS software which was developed by the National Library of Australia, and the collections were "hosted by an external agency" (Pennock, 2013, p. 27). In 2008, UKWAC moved to an in-house operation using the Web Curator Tool (WCT) workflow management tool which was collaboratively developed by the British Library and the National Library of New Zealand through an International Internet Preservation Consortium (IIPC) funded project (Pennock, 2013, p. 27; Web Curator Tool, n.d., History). However, with the implementation of NPLD regulations in 2013, this brought about a major transformation in web archiving for the UK legal deposit libraries, "necessitating new workflows to deal with the selection, annotation and curation of content harvested both as part of the broad domain crawls and as part of more frequent and targeted crawling activity" (UK Web Archive, n.d., W3ACT User Guide). The WCT was not scalable, so the Annotation Curation Tool (W3ACT) was developed "to meet the requirements of subject specialists wishing to curate web content harvested under the Legislation" (UK Web Archive, n.d., W3ACT User Guide, also Jackson, 2016a). For example, W3ACT allows "users to perform numerous curatorial tasks including the assignation of metadata and crawl schedules to web content, quality assurance and the ability to request permission for open access to selected websites" (UK Web Archive, n.d., W3ACT User Guide). Additionally, the change in curation tool, also necessitated a change in crawling software from HTTrack to Heritrix (Pennock, 2013, p. 27).

The UK Web Archive personnel are keen users of open-source tools and in doing so, they contribute back to the web archiving community (Jackson, 2022a; UK Web Archive, 2018). They also offer research support to PhD students wishing to use their collections. They have also collaborated with academic communities to develop tools for users such as SHINE. SHINE is a prototype of a potential research tool that can be used to access and analyse web archive data. It was developed as part of the Big UK Domain Data for the Arts and Humanities project funded by the UK Arts and Humanities Research Council. The data that underpins this service was acquired by JISC from the Internet Archive and includes all .uk websites in the Internet Archive web collection crawled from 1996 to April 2013, when NPLD came into effect (UK Web Archive, n.d., JISC UK Web Domain Dataset). The JISC UK Web Domain Dataset is available for use through the UK Web Archive website and listed in the British Library Shared Research Repository. Users can also search on SHINE either using a URL or keywords. The search results can then be further filtered by using predefined facets. In

addition, trend graphs can be generated by using keywords and a time range between 1996 and 2013. Clicking on a single point in a trend graph will generate a sample of 100 resources that reference that keyword and link out to the Internet Archive (UK Web Archive, n.d., SHINE; Byrne, 2019).[36] The JISC UK Web Domain Dataset (1996-2013) also offers four derived datasets which can be reused by researchers for big data analysis (UK Web Archive, n.d., JISC UK Web Domain Dataset). Furthermore, the dataset offers an example of how 'some' UK digital heritage was salvaged for the years before the legislation allowed for selective web archiving from 2004, and for the crawling of the UK national web domain from 2013.

The UK Web Archive offers a selection of collections, with content both open access and onsite access, which may be representative of the NI web space. A few collections to note here are the UK General Election collections which cover the NI web space, and the Gender Equality collection which has a bodily autonomy subsection that covered the 8th Amendment Referendum and the Now for NI campaign. NI is further represented in sporting collections although there are some gaps. The News collection has a wide variety of NI publications. The Easter Rising 1916/2016 collection is also relevant for both NI and ROI. The UK Web Archive and PRONI have also collaborated on several curated web archive collections to offer a NI perspective on topics and events including Brexit, the 2019 UK General Election, Covid-19, UEFA Women's Euro England 2022, and these are available through the UK Web Archive (Murchan, 2020b). All published collections are visible through the UK Web Archive website on the Topics and Themes page (see Figure 5.5).[37]

End users should be aware that the records in the collections contain a mixed model in terms of access; some archived websites being open access and some only available onsite at legal deposit libraries. Nonetheless, it is useful to have collections arranged by topic and theme. This structure enables researchers to assess whether or not the holdings might warrant a trip to the TCD library to view the websites onsite at a library terminal. However, as mentioned previously in section 2.3, onsite access to the UK national web domain collections presents multiple challenges often due to the restrictive nature of the UK legal deposit legislation (NPLD). Indeed, Gooding et al. (2021) report that "many researchers have publicly questioned whether the restrictive access protocols for NPLD are in fact a barrier to the usage of electronic publications" (p. 1155). Furthermore, they suggest that the NPLD protocols are not in line with current trends in digital user expectations and information seeking behaviours (Gooding et al., 2019, p. 21). It should also be noted that by the time the

[36] UK Web Archive SHINE, https://www.webarchive.org.uk/shine

[37] UKWA Topics and Themes, https://www.webarchive.org.uk/en/ukwa/category/

2013 NPLD Regulations had come into effect, publishing and communications technologies had rapidly advanced in the meantime (Arnold-Stratford & Ovenden, 2020, p. 5).[38]

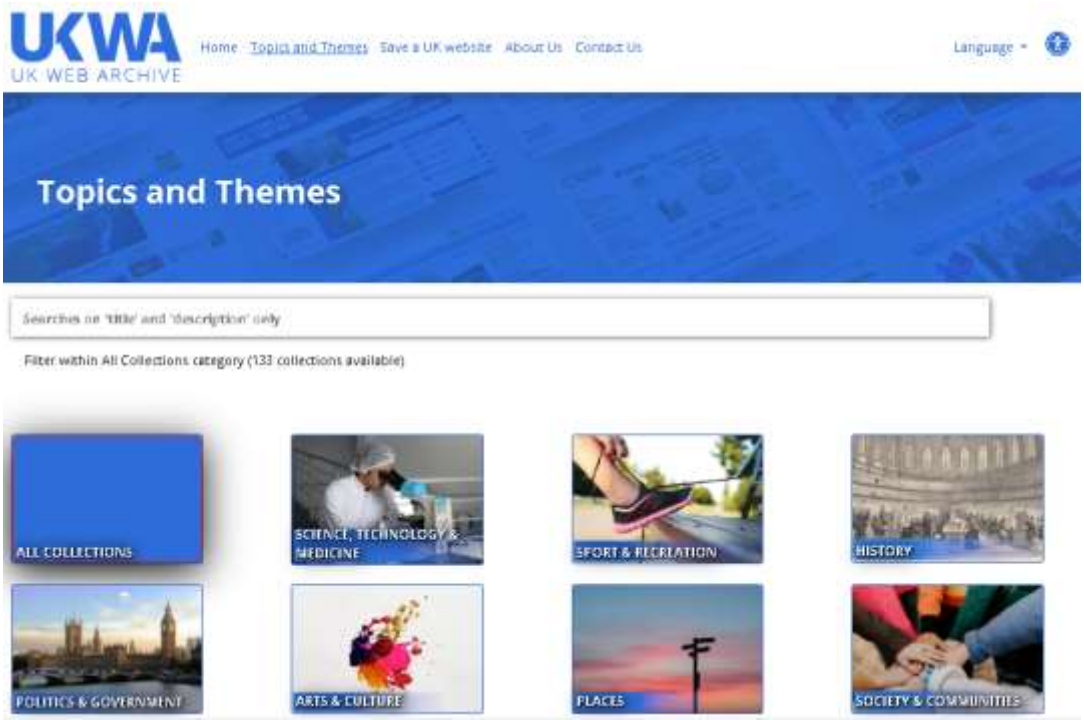

Figure 5.5 Screenshot of UKWA web page for Topics and Themes, taken on 2022-10-16

5.3.3 NI In Brief

While there are resource and legislative limitations, concrete efforts are being made to provide a balanced approach towards the collection and preservation of the NI web space. Firstly, the UK Web Archive captures and preserves websites from the NI web space, through a selective collecting approach and through an annual domain crawl of the UK web space. This content is accessible onsite in UK legal deposit libraries, inclusive of the Library of TCD in Dublin. As outlined previously, only a small percentage of this collection has been made open access with consent from the content owners. Secondly, a two-fold approach by PRONI provides a publicly accessible selective web archive collection through (i) the collection of websites of government, public bodies etc., with notifications of the intent to collect, and provisions of a takedown policy, and (ii) a permissions-based approach for privately funded websites. And thirdly, through a collaborative effort, the PRONI Web Archive and the UK Web archive are developing curated collections. Access to the collections differ, with PRONI

[38] Legal Deposit Libraries Act 2003 (Commencement) Order 2004, https://www.legislation.gov.uk/uksi/2004/130/introduction/made

Web Archive being open access, and the UK Web Archive being a mix of both open and onsite access. However, as discussed earlier, onsite access presents multiple challenges for end users due to the restrictive nature of current UK legal deposit legislation, while onsite access is economically non-viable for many users.

5.4 ROI Web Space

In terms of a legal deposit strategy, it would make sense for the ROI to organise a collection development strategy, similar to the UK/NI model, for the capture and preservation of the ROI web space. In this way future generations will have access to a more representative landscape of Irish digital heritage on the web. However, there would need to be a more realistic approach towards the provision of access, considering the challenges that are currently being presented due to the restrictive nature of UK legal deposit legislation, which are in essence barriers for innovation. The JISC UK Web Domain Dataset (1996-2013) also offers a model for the salvage of 'some' digital heritage for the ROI's .ie domain from 1996 to the current day.

In addition, while the UK model on demarcating a national web domain would be an equally good example for demarcating the Irish web domain, there is also the need to consider that there are "significant gaps in the heritage acquired as websites on 'non-UK' top level domain names, such as .com, are not automatically identified" (Bingham & Byrne, 2021, p. 2). To overcome this, it would be useful to also look at the Icelandic model which "contains all web sites hosted on the Icelandic domain .is" as well as many websites "hosted elsewhere that are in Icelandic or refer directly to matters of interest to Iceland" (IIPC, n.d., Landsbókasafn Íslands). The Icelandic Web Archive is also open access except for websites where the user must pay for access and when the content owners have requested that access to the archived version of the content is blocked (IIPC, n.d., Landsbókasafn Íslands).

Thus, the ROI should evaluate the demarcation of its national web domain to ensure such gaps are minimised. However, at the current time, the inclusion of a national web domain archive is not part of ROI legal deposit legislation, nor does there seem to be any efforts (in the public domain at least) towards the salvage of ROI digital heritage from any other web archive. Moreover, for the most part, the parliamentary discussions regarding the establishment of a national web domain archive are solely focused on the ROI's .ie domain (Figure 5.11).

While many western societies undertook a review of their copyright, heritage, and legal deposit laws from the 1990s, due to advances in publishing and communication, and many

have implemented reform to account for legal deposit of non-print materials, including the development of national web archiving programmes through a selective based approach, and routine national domain web archiving, the ROI has trailed behind Canada, New Zealand, and much of Europe (Conul, 2012, p. 14).

5.4.1 NLI Web Archive

As mentioned before, the NLI is the official library of record in the ROI and is responsible for collecting for, and on behalf of the Irish state (NLI, Collection Development Policy 2022-2026). The NLI commenced a Born Digital Programme in 2011 to

> identify the role of the National Library of Ireland in relation to the collection of born-digital material and to identify, collect and make accessible born-digital material as part of day-to-day collection development activities (NLI Annual Report 2011, p. 8).

As part of this, following public procurement processes the NLI teamed up with the IMF subscription-based web archiving service in Europe for a selective (permissions-based) web archiving initiative for the 2011 General Election. The project entailed the capturing of 100 websites to include: candidate's sites, political blogs, news media, and some official government sites. Later the same year, web content was collected for the 2011 Irish Presidential Election (NLI Annual Report, 2011, p. 8). Since then, the NLI has continued to operate a small-scale selective web archiving programme for the capture of Irish social, cultural, political, governmental, and commemorative heritage on the web. The collections are openly accessible via the NLI Web Archive. Following a procurement process in early 2018, the NLI commenced using the Internet Archive's Archive-It subscription service and migrated its holdings from the IMF platform to their current location on the Archive-It platform. Figure 5.6 offers a screenshot of the initial public NLI Web Archive via the IMF interface from 2015. Also worth mentioning is the fact that the migration of the NLI Web Archive from one platform to another would have caused some disruption for end users who were using the initial IMF resource for research at the time (Aturban, 2019a). Ultimately all the URLs linking to the data in the original NLI Web Archive through the IMF platform became invalid (http://collection.europarchive.org/nli/), while the Archive-IT web archive offers a new URL (https://archive-it.org/home/nli).

As of September 2022, the selective NLI Web Archive contains 75 collections with 3,105 websites overall which cover a diverse range of topics such as Irish History, LGBTI+, Irish Entertainment, Ageing in Ireland, Coastal & Island Life, Irish Music, Higher Education, Agriculture, Horticulture & Food, and Housing & Property, to name but a few. One can

browse the collections alphabetically or browse an alphabetical listing of the individual websites (Figure 5.7 & Figure 5.8). The resource can also be searched by inputting a known URL, and through a full-text search. Both the pre-2018 archived websites (migrated from the IMF platform) and the post-2018 archived websites are organised by collection on the Archive-It platform with applicable metadata which is very beneficial for end users (see for example the General Election 2011 collection).[39] Indeed, the application of metadata is a noteworthy undertaking by the NLI considering it is a resource intensive undertaking (Costa, 2021, p. 72), coupled with the fact that the NLI is already "constrained primarily by limited staff numbers" (Collins, 2018, p. 180). Equally useful for understanding some of these earlier collections are the NLI selective web collection development policies (Figure 5.9; NLI, 2022, Selective Web Archive Collections).

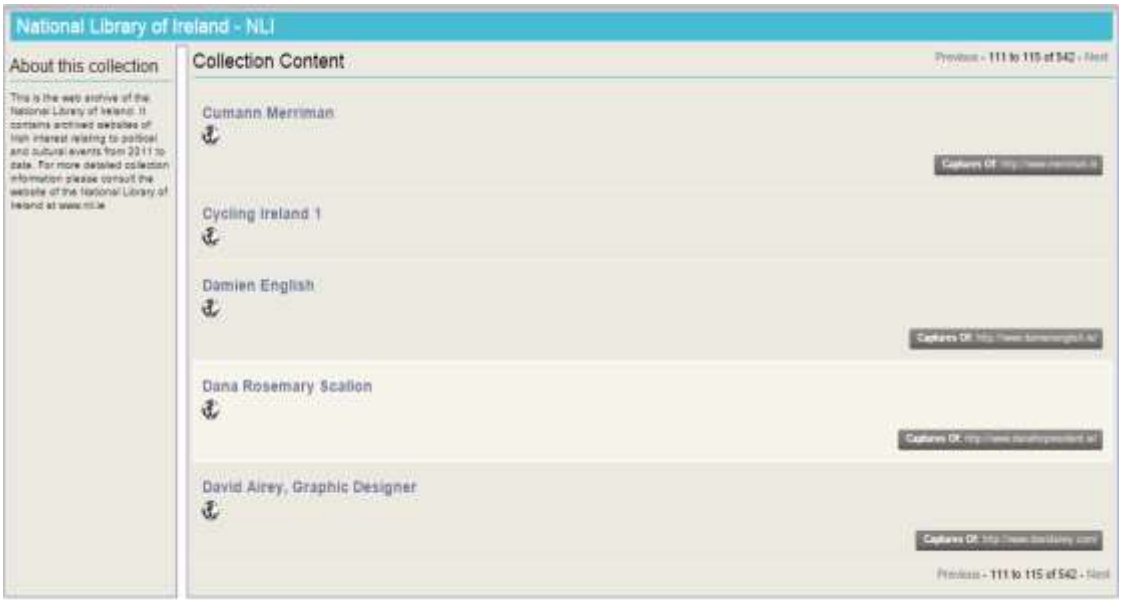

Figure 5.6: Screenshot of NLI Web Archive user interface on the Internet Memory Foundation platform, taken in June 2015 (personal archive)

[39] NLI, Archive-It - General Election 2011, https://archive-it.org/collections/19959

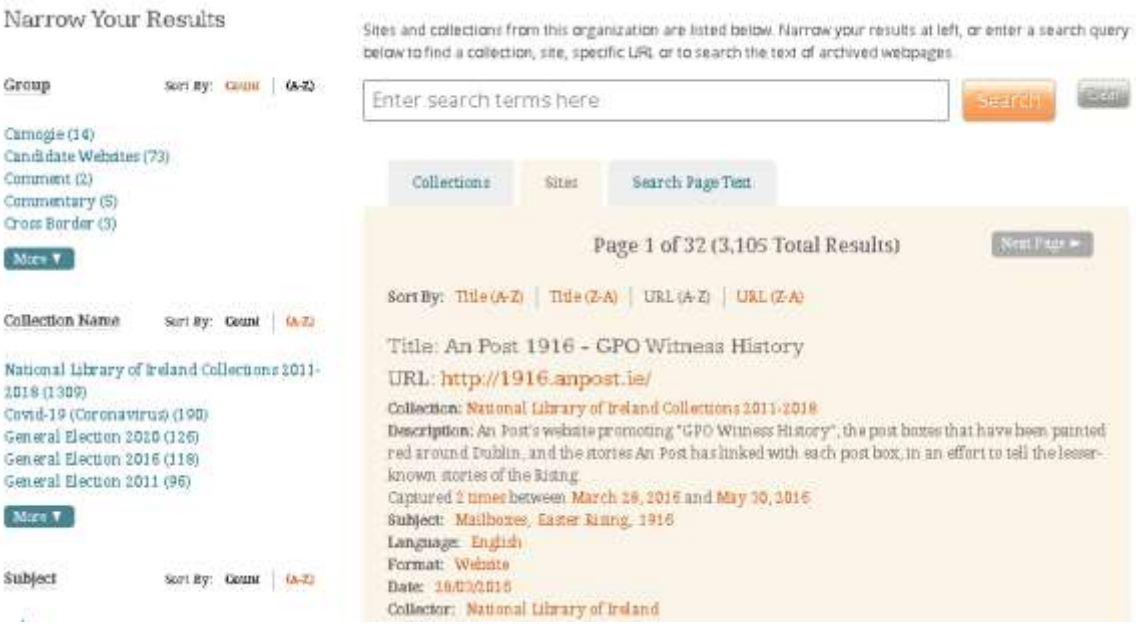

Figure 5.7: Screenshot of NLI Web Archive interface on the Archive-It platform showing a total of 3,105 websites in their collections, taken on 2022-09-28

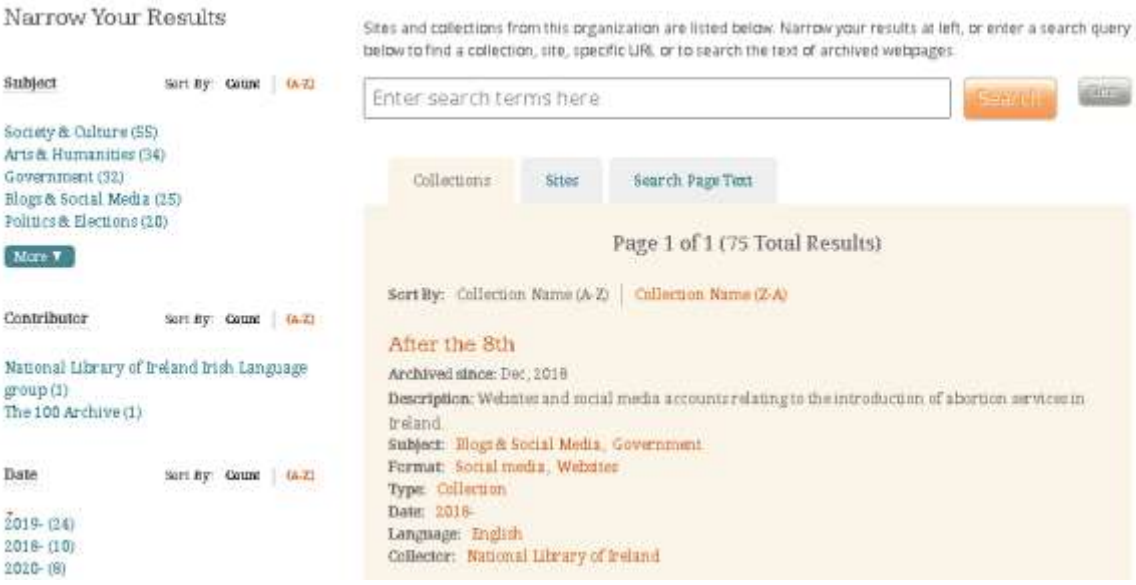

Figure 5.8: Screenshot of NLI Web Archive interface on the Archive-It platform showing a total of 75 collections, taken on 2022-09-28

Figure 5.9: Screenshot of NLI website with collection lists for the selective web archive collections pre-2018, taken 2022-09-09

Further to this, following public procurement processes the NLI collaborated with the Internet Archive web archiving service to capture the Irish ccTLD web domain (.ie) in 2007 and again in 2017 (NLI, 2017; NLI, 2022, Irish Domain Web Archive). In 2017, *The Irish Times* online speculated that the domain crawls would become available for access onsite in the NLI reading rooms (Taylor, 2017a). However, this has yet to happen, and there have been no further domain crawls conducted since 2017. The issue here for both the archiving of the Irish web domain (.ie), and the provision of access to the domain content that is already captured, is due to the current state of Irish legislation on copyright and legal deposit.

While the NLI is a legal deposit library, digital legal deposit legislation was not enacted in Ireland at the time the 2007 or 2017 domain crawls were conducted (O'Dell, 2018; Collins, Joint Committee on Tourism, Culture, Arts, Sport and Media, 13 October 2021). Moreover, while digital legal deposit legislation came into force in December 2019 through the Copyright and Other Intellectual Property Law Provisions Act 2019 (hereafter, COIPLPA, 2019), it did not include a clause for crawling the Irish national web domain. Thus, the 2007 and 2017 domain crawls remain inaccessible to both researchers and the public alike, and the NLI is prohibited from doing any further domain crawls. However, COIPLPA (2019) does contain a clause as follows:

> Within twelve months of the enactment of this Act the Government shall bring forward a report on the feasibility of establishing a digital legal deposit scheme

to serve as a web archive for .ie domain contents and advise on steps taken towards that goal (Section 108: COIPLPA, 2019; Figure 5.10).

Regrettably, to date, the ROI government has failed to deliver a feasibility report and thus failed to incorporate a domain web archive as part of copyright and legal deposit legislation. Consequently, the NLI continues to be prohibited from doing any further crawls, while the country continues to suffer from mass losses of Irish digital heritage. For more understanding of these issues, we offer an overview of copyright and legal deposit legislation in the ROI.

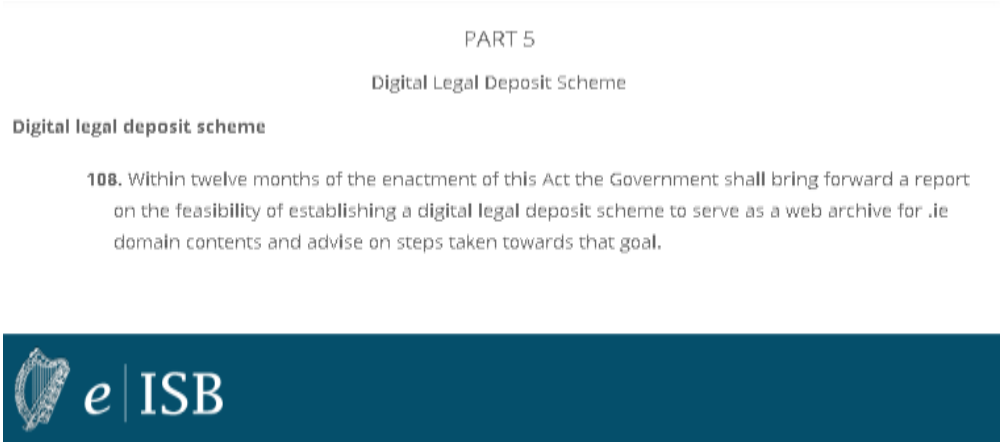

Figure 5.10: Screenshot of Section 108 in the Copyright and Other Intellectual Property Law Provisions Act 2019, taken on 2022-10-07

Following partition, a legal deposit scheme for print publications was introduced in 1927 through the Industrial and Commercial Property (Protection) Act, 1927 (ICPPA, 1927).[40] This established the relationship between copyright and legal deposit in the same legislation. The ICPPA (1927) also formalised the NLI as a legal deposit library alongside the Library of TCD, the British Museum (now the British Library), and the three constituent college libraries of the National University of Ireland (at the time), being University College, Dublin, University College, Cork, and University College, Galway (Section 178, Ireland, ICPPA, 1927). The next change in the law which applied to legal deposit was in Section 56 of the Copyright Act, 1963, whereby it included the library of St. Patrick's College, Maynooth (later Maynooth University) as a designated legal deposit library.[41] Further amendments to legal deposit legislation saw the introduction of two new legal deposit libraries through Section 7 of the

[40] Industrial and Commercial Property (Protection) Act, 1927, https://www.bailii.org/ie/legis/num_act/1927/0016.html#zza16y1927

[41] Copyright Act, 1963, https://www.irishstatutebook.ie/eli/1963/act/10/enacted/en/html

University of Limerick Act, 1989, and Section 6 of the Dublin City University Act, 1989.[42] Currently, the ROI legal deposit institutions include: the National Library of Ireland, the Library of Trinity College Dublin, the British Library, and the libraries of Dublin City University, National University of Ireland, Galway, Maynooth University, University College Dublin, University College Cork, and University of Limerick. In addition, Irish publishers may also be obliged to deposit publications, on request by the Bodleian Library, University of Oxford, the Cambridge University Library, the National Library of Scotland, and the National Library of Wales (Section 198: Copyright and Related Rights Act, 2000; National Library of Ireland, 2014).[43]

While there were subsequent amendments to ROI copyright law during the 1990s, the amendments did not affect legal deposit per se. Ireland was also a member of the EEC since 1973 and was obliged to keep in line with their directives. For example, the European Communities (Legal Protection of Computer Programs) Regulations, 1993 was introduced by the ROI government to accommodate copyright for computer programs and programmers in compliance with EEC Council Directive 1991/250/EEC.[44] And like the UK, the ROI also had to comply with the EEC Council Directive No. 93/98/EEC, to harmonise copyright across member states, and thus, the Irish government introduced the European Communities (Term of Protection of Copyright) Regulations, 1995.[45]

The Copyright and Related Rights Act, 2000 (CRRA) was also established in some parts to comply with EC directive 2001/29/EC – Harmonisation of certain aspects of copyright and

[42] University of Limerick Act, 1989,
https://www.irishstatutebook.ie/eli/1989/act/14/enacted/en/html;
Dublin City University Act, 1989,
https://www.irishstatutebook.ie/eli/1989/act/15/enacted/en/html

[43] Copyright and Related Rights Act, 2000, Section 198,
https://www.irishstatutebook.ie/eli/2000/act/28/section/198/enacted/en/html

[44] European Communities (Legal Protection of Computer Programs) Regulations, 1993,
https://www.irishstatutebook.ie/eli/1993/si/26/made/en/print; EEC Council Directive 91/250/EEC- Legal protection of computer programs, https://op.europa.eu/en/publication-detail/-/publication/92d68447-ea9a-4554-9540-de517984c310/language-en

[45] European Communities (Term of Protection of Copyright) Regulations, 1995,
https://www.irishstatutebook.ie/eli/1995/si/158/made/en/print;
EEC Council Directive 93/98/EEC, http://data.europa.eu/eli/dir/1993/98/oj/eng

related rights in the information society.[46] Furthermore, section 199 of the CRRA recognises the need for the inclusion of non-print materials such as:

> any engraving, photograph, text of a play, cinematograph film, microfilm, video recording, sound recording, record, diskette, magnetic tape, compact disc, or other thing on or in which works or information or the representations thereof is written, recorded, stored or reproduced but does not include local records or local archives within the meaning, in each case, of section 65 of the Local Government Act, 1994, or books within the meaning of section 198 of the Copyright and Related Rights Act, 2000 (Section 199: CRRA, 2000).

Section 198 of the CRRA relates to the legal deposit of books, defining a book to include:

> every part or division of a book, pamphlet, sheet of letterpress, sheet of music, map, plan, chart or table separately published, but shall not include any second or subsequent edition of a book unless such edition contains additions or alterations either in the letterpress or in the maps, plans, prints or other engravings belonging thereto (Section 198, CRRA, 2000).

The CRRA also makes an acknowledgement for the allowance of the deposit of books in electronic form in Section 198. As outlined below, however, the CRRA did not consider the fact that some publications are born digital only, with no print counterpart, such as internet journals and publications, e-zines, or web pages.

> (11) Where a copy of a book requested under subsection (1) is delivered in a form other than an electronic form, the Board or other authorities referred to in subsection (1) may request, in addition to that copy, a copy in an electronic form readable by means of an electronic retrieval system and on such request being made a copy in electronic form shall be delivered by the publisher to the Board or authority concerned.
>
> (12) For the purposes of this section, "publication", in relation to a book—
>
> a) means the issue of copies to the public, and
> b) includes its making available to the public by means of an electronic retrieval system,
>
> and related expressions shall be construed accordingly (Section 198: CRRA, 2000).

[46] Copyright and Related Rights Act, 2000, https://www.irishstatutebook.ie/eli/2000/act/28/enacted/en/html;
EC Council Directive 2001/29 - Harmonisation of certain aspects of copyright and related rights in the information society, http://data.europa.eu/eli/dir/2001/29/oj/eng

Recognising the need for measures to preserve Irish digital heritage, both the Library of TCD and the NLI, as nominated legal deposit libraries in the ROI, instigated different schemes to accommodate the collection of electronic and born digital publications. In the NLI Collection Development Strategy 2009-2014, the NLI asserts that: "If the challenge of collecting Ireland's online presence is not addressed, it can be argued that the Library is not respecting its remit as a national memory institution" (p. 9). In a bid to provide some solutions, the NLI set up the Born Digital Programme in 2011, for the purpose of identifying and collecting born digital content and implementing web archiving practices as a regular activity of the library (NLI Annual Report 2011, p. 8; NLI Annual Report 2012, p. 13). The Library of TCD set up a voluntary electronic deposit scheme in Ireland through their resource edepositIreland, as a "self-deposit service [...] open to all publishers in Ireland, be they individuals, local groups, publishing houses or organisations, who wish to share their publications with the world" (Library of TCD, n.d., edepositIreland). Also, as mentioned earlier the Library of TCD provides access to the UK Web Archive legal deposit domain web archive, which is accessible onsite through the library pc terminals.

5.4.2 Debating the Issue

It was not until 2011 that we witnessed a more serious investment by the ROI government (a coalition of the Fine Gael and Labour Party) to address both copyright and legal deposit due to advances in the web and the internet (O'Dell, 2013). In May 2011, Richard Bruton (Fine Gael), who was the Minister for Jobs, Enterprise, and Innovation at the time, established the Copyright Review Committee (CRC) to:

> Examine the present national copyright legislation and identify any areas that are perceived to create barriers to innovation [and] Identify solutions for removing these barriers and make recommendations as to how these solutions might be implemented through changes to national legislation (CRC, 2012, p. 1; Department of Jobs, Enterprise, and Innovation, 2012, Consultation).

The CRC held a public meeting in July 2011 in Dublin and from there solicited more than 100 written submissions regarding copyright review overall. The submissions were at one time available to view on the website for the Department of Jobs, Enterprise and Innovation, but the website is no longer available on the live web, as the department was reformulated twice since then, which unsurprisingly meant new URLs for their websites (see Table 5.1). Therefore, one needs to use the NLI Web Archive to view the different types of stakeholders who made submissions (Figure 5.11). It is also worth discussing the issues with ROI government department websites a little further.

Figure 5.11: Screenshot of the website for the Department of Jobs, Enterprise and Innovation, with the submissions received by the Copyright Review Committee, captured in the NLI Web Archive (Timestamp: 2012-06-13 23:06:22) [47]

Table 5.1 shows how one ROI government department was reformulated five times from 1997 to 2020, resulting in five different titled websites, of which the latest website only lists news/media items as far back as 2016, meaning that news/media items from 1997-2015 may only be found in a web archive, if at all. While the NLI has also been capturing ROI government department websites since 2011, scholars will need to rely heavily on the Wayback Machine for Irish web content created from the mid-1990s to 2011, in a bid to find fragments of what was at one time public information provided by ROI government departments on their websites (Figure 5.12). However, the Wayback Machine may only hold surface pages of an individual departmental website, and not have crawled the multitude of internal hyperlinks for departmental news or publication items for example. Moreover, there are undoubtedly a multitude of academic publications and government publications which contain URL references, linking to these departmental websites over the years which are no longer valid.

[47] Department of Jobs, Enterprise and Innovation, Submissions Received by the Copyright Review Committee, NLI Web Archive, 2012, https://wayback.archive-it.org/org-1444/20120613230622/http://www.djei.ie/science/ipr/crc_submissions2.htm

Department of Enterprise, Trade and Employment
Timestamp: 1999-10-04
(www.entemp.ie)

Department of Enterprise, Trade and Employment
Timestamp: 2002-05-29
(www.entemp.ie)

Department of Enterprise, Trade and Innovation
Timestamp: 2010-05-07
(www.deti.ie)

Department of Jobs, Enterprise, and Innovation
Timestamp: 2011-06-06
(www.djei.ie)

Department of Jobs, Enterprise and Innovation
Timestamp: 2012-06-10
(www.enterprise.gov.ie)

Department of Business, Enterprise and Innovation
Timestamp: 2017-09-05
(www.dbei.gov.ie)

Figure 5.12: The changing nature of ROI Government Department websites and URLs is revealed by examining the first captures of their websites in the Wayback Machine

Table 5.1: Renaming of Department of Jobs, Enterprise and Innovation before and after formulation (Sources: Wikipedia, 2005+)

Dates	Department name	Website URL	URL Status notes	Link Rot
July 1997	Renamed as the Department of Enterprise, Trade and Employment	http://entemp.ie	Redirects to https://www.iva-advice.co	YES
May 2010	Renamed as the Department of	http://www.deti.ie	Page not Found	YES

	Enterprise, Trade and Innovation			
June 2011	Renamed as the Department of Jobs, Enterprise and Innovation	http://www.djei.ie	Redirects to new website, https://enterprise.gov.ie/djei/en/	YES
Sept. 2017	Renamed as the Department of Business, Enterprise and Innovation	https://dbei.gov.ie/en	Redirects to new website, https://enterprise.gov.ie/djei/en/	YES
Nov. 2020	Renamed as the Department of Enterprise, Trade and Employment	https://enterprise.gov.ie/djei/en/	Still there, pending migration to the central government website	

Getting back to the CRC and copyright review, from an analysis of the written submissions on copyright review, the CRC published a consultation paper in 2012 which proposed amendments to copyright legislation and solicited further consultations and feedback for the proposals (CRC, 2012, Copyright and Innovation: A Consultation Paper). Of interest in the consultation paper is the classifications of the submissions into the categories of "(i) rights-holders; (ii) collecting societies; (iii) intermediaries; (iv) users; (v) entrepreneurs; and (vi) heritage institutions" (O'Dell, 2012; CRC, 2012, pp. 9–10). The Consultation Paper breaks these categories as outlined verbatim below, which offers us an overview of the stakeholders with an interest in Irish copyright review.

- rights-holders: this category includes the people who create the copyright work, and as well as their publishers, music labels, movie studios, broadcasters and so on,
- collecting societies: this category includes societies which grant licences of copyrighted works and collect copyright royalties for distribution back to the rights-holders,
- intermediaries: this category includes internet service providers, online search engines, social networks, and trading sites,
- users: this category includes the consumers, purchasers and users of copyright works,
- entrepreneurs: this category includes online start-ups,
- heritage institutions: this category includes libraries, archives, galleries, museums, schools, universities and other educational establishments, and the like (Copyright and Innovation: A Consultation Paper, pp. 9–10).

Following further analysis, the CRC then published the Modernising Copyright report in October 2013, which offered modern solutions to Ireland's outdated copyright laws.

Of particular interest, the Modernising Copyright report makes recommendations for the introduction of digital legal deposit to current legal deposit institutions, and further to this, that such institutions should be permitted to "make copies of our online digital heritage by reproducing any work that is made available in the State through the internet" (CRC, 2013, p. 5). In this context, the report clarifies the meaning of a work on the internet to be:

> a work shall have been made available in the State through the internet where (a) it is made available to the public either from a website with a domain name which relates to the State or to a place within the State, or by similar or related means, or (b) it is made available to the public either by a person any of whose activities relating to the creation or the publication of the digital publication takes place within the State, or by a person with similar or related connections to the State (CRC, 2013, p. 153).

The report further advocates for the "formation of a Copyright Council of Ireland, as an independent self-funding organisation, created by the Irish copyright community, recognised by the Minister, and supported and underpinned by clear legislative structures provided" (CRC, 2013, p. 9). The purpose of which would serve to

> ensure the protection of copyright and the general public interest as well as encouraging innovation; and it should have a broad subscribing membership and a Board drawn widely from the Irish copyright community. It should provide education and advice on copyright issues, advocate both nationally and internationally for developments in copyright policies or procedures, and work towards solutions on difficult copyright issues. It should be able to establish a Digital Copyright Exchange (to expand and simplify the collective administration of 10 copyrights and licences), a voluntary alternative dispute resolution service (to meet the need for an expeditious dispute resolution service outside the court system), and an Irish Orphan Works Licensing Agency (to provide a solution to the problem of orphan works) (CRC, 2013, p. 9).

Indeed, a Copyright Council of Ireland would also make sense in terms of the need for a continual evaluation of legal deposit legislation in line with the fragility of born digital heritage and the technological advances in publishing and communication technologies. Nonetheless, the Irish government was slow to embrace the zeitgeist of the recommendations, despite real-time concerns for the loss of Irish digital heritage on the web in the meantime.

Copyright amendments were imminent, due to the advances in the web and the internet in Ireland from early 2000 (Sterne, 2015+), and the prevalence of copyright infringements online from mid-2000 onwards (e.g., music, film, photography, etc.) (O'Dell, 2013; Morris, 2019). There was also a need for changes to copyright to facilitate disability requirements and educational needs, such as allowing for the use of copyrighted multimedia on a classroom whiteboard, and for allowing the modification of books to meet the needs of individuals with disabilities in line with the international Marrakesh Treaty to Facilitate Access to Published Works for Persons who are Blind, Visually Impaired, or otherwise Print Disabled (Department of Enterprise, Trade and Employment, 2016). However, the inclusion of "digital" with regards to legal deposit could not be assumed. As O'Dell (2016) points out, when the government finally announced the drafting of the Copyright and Related Rights (Miscellaneous Provisions) Bill, 2016, they were initially opting for the incorporation of a digital legal deposit scheme "on a voluntary basis" (O'Dell, 2016; Department of Enterprise, Trade and Employment, General Scheme, 2016).

Another consultation was launched in April 2017, by the reformulated Department of Arts, Heritage, Regional, Rural and Gaeltacht Affairs (previously Department of Arts, Heritage and the Gaeltacht) in consultation with the NLI. Titled 'Consultation on the Legal Deposit of published digital material in the 21st century in the context of Copyright legislation', it was aimed "at gathering stakeholder views in regard to whether or not the policy in relation to Legal Deposit should include the collecting, preserving and making available of all contemporary publication formats, including online digital formats such as websites." The consultation requested opinions from "the library and archives community, publishers and members of the public in the context of the review of the Copyright and Related Rights Act" (Department of Arts, Heritage, Regional, Rural and Gaeltacht Affairs, 2017).

In August 2017, the Department of Arts, Heritage, Regional, Rural and Gaeltacht Affairs was reformulated to become the Department of Culture, Heritage, and the Gaeltacht, and would retain responsibility for addressing legal deposit. In December 2017, the reformulated department issued a response to the 42 submissions it received for the consultation process. Regarding the extension of legal deposit to include digital formats inclusive of websites, the Department responded with the following statement:

> 93% of responses to this question were strongly supportive, with respondents highlighting the ephemeral nature of online digital material, and the huge threat of its loss, unless institutions such as the National Library of Ireland operating in the cultural heritage area are legally mandated to preserve it. Many respondents also referred to the fact that the history of the 21st century is recorded online, and how the loss of this online information will lead to the loss of 'significant

> national documentation' as well as the loss to researchers of the outputs of research (Department of Culture, Heritage, and the Gaeltacht, 2017).

Thus, there were recommendations by the *Modernising Copyright* (2013) report for the establishment of an Irish web domain archive, and 93% of the submissions to the Consultation on Legal Deposit in 2017, were also in support of the establishment of an Irish web domain archive.

The following year saw the introduction of the Copyright and Other Intellectual Property Law Provisions Bill 2018 (as initiated) which was brought before Dáil Éireann in March 2018,

> to amend the Copyright and Related Rights Act 2000 to take account of certain recommendations for amendments to that Act contained in the Report of the Copyright Review Committee entitled 'Modernising Copyright' published by that Committee in October 2013 and also to take account of certain exceptions to copyright permitted by Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (Copyright and Other Intellectual Property Law Provisions Bill 2018 (as initiated), 2018).

Nonetheless, while the Bill (as initiated) "extended the copyright deposit regime to ebooks", it did not "provide for the harvesting of the .ie domain" (O'Dell, 2018). Fianna Fáil then put forward the following amendment which was approved by Dáil Éireann.

> Within twelve months of the enactment of this Bill the Government shall bring forward a report on the feasibility of establishing a digital legal deposit scheme to serve as a web archive for .ie domain contents and advise on steps taken towards that goal (amendment cited in O'Dell, 2018).

O'Dell suggests that this was at least progress, "even if it amounted to making haste slowly" (2018). However, at the Seanad Éireann committee stage, Senator Fintan Warfield (Sinn Féin) insisted that

> The time for examining feasibility has long passed. It should have been done as soon as it was recommended [in 2013]. The only way we can have certainty in respect of this issue is to provide for it in law through this Bill and I respectfully encourage the Government to do so (Fintan Warfield, Seanad Éireann, Copyright and Other Intellectual Property Law Provisions Bill 2018: Committee Stage, 03 October 2018).

Thus, Warfield proposed an amendment as outlined below.

> (c) For the purposes of this subsection, a work shall have been made available in the State through the internet where—

> (i) it is made available to the public either from a website with a domain name which relates to the State or to a place within the State, or by similar or related means, or
>
> (ii) it is made available to the public either by a person any of whose activities relating to the creation or the publication of the digital publication takes place within the State, or by a person with similar or related connections to the State (Fintan Warfield, Seanad Éireann, Copyright and Other Intellectual Property Law Provisions Bill 2018: Committee Stage, 03 October 2018).

However, the Minister of State for Training, Skills, Innovation, Research and Development, John Halligan (Independent) had responsibility for carrying the Bill on behalf of the Fine Gael and Independent coalition government and objected to Warfield's amendment as outlined below.

> Providing for a full digital deposit system that would facilitate capturing the web is not simply a matter of changing copyright legislation. It is a significant national project that requires multi-institutional collaboration, significant resources and Skillsnet [*sic*] for capturing and preserving Ireland's digital record, according to my advice. I reiterate that this is a matter for the Minister for Culture, Heritage and the Gaeltacht who has responsibility for policy in this area. My Department and the Department of Culture, Heritage and the Gaeltacht have actively worked together on that matter for some time and we will continue to do so until the robust regulatory framework is developed. We will facilitate the necessary corresponding legislation amendments in due course [...]
>
> This work, however, is not yet sufficiently progressed for any technical amendments to copyright law. As that is the final aspect of the project, now that all the necessary due diligence has been done, Government mechanisms have been established and funding has been agreed with the Minister for Public Expenditure and Reform, it is not possible for amendments to copyright law to be progressed in isolation from Government approval for the project as a whole [...]
>
> [A] new section stipulates that a report be published within 12 months of the Bill being enacted. This was accepted by all parties and viewed as a pragmatic way to advance the project while allowing time for the necessary work to take place in the Department of Culture, Heritage and the Gaeltacht, in co-operation with my Department and the Department of Public Expenditure and Reform. The House can rest assured that both Departments are actively engaged in advancing the proposal and the report will be prepared within 12 months, as specified in

> the Bill (John Halligan, Seanad Éireann, Copyright and Other Intellectual Property Law Provisions Bill 2018 – Committee Stage, 03 October 2018).

Nonetheless, Senator Warfield "pressed it to a vote" (O'Dell, 2018). According to O'Dell (2018): "On the electronic vote, there was a tie – Tá (yes) 18; Níl (no) 18 – and the amendment was defeated on the casting vote of the Leas Cathaoirleach (Deputy Speaker)." However, Warfield "called for a walk-through vote, and the amendment was [carried] – Tá (yes) 19; Níl (no) 17" (O'Dell, 2018). Nonetheless, when it went to Report Stage in Dáil Éireann in May 2019, Minister Halligan

> unapologetically restated his objections that there were issues with other government departments and public institutions, and that it would have significant resource implications, and he put down an amendment to reverse Senator Warfield's earlier successful amendment (O'Dell, 2019).

Finally, a digital legal deposit scheme was formally organised in the ROI through the Copyright and Other Intellectual Property Law Provisions Act 2019 (COIPLPA, 2019).[48] The Act allows for the collection of e-books and journals on the internet, but it does not allow for the archiving of the Irish national web domain (Ryan et al., 2022). It does, however, provide a clause to "bring forward a report on the feasibility of establishing a digital legal deposit scheme to serve as a web archive for .ie domain contents and advise on steps taken towards that goal" within twelve months of the Act coming into force in December 2019 (Figure 5.10). Yet, as of October 2022, a feasibility report has still not been produced.

Certainly, a feasibility study may have been disrupted due to the onset of the COVID-19 pandemic in Ireland in March 2020 (O'Dell, 2020). However, as the country gets back to normal, the pandemic should no longer be a reason for the holdup. ROI Parliamentary questions, and committee debates, also provide some indications as to the hold-up in adopting routine domain web archiving as a necessary component of a modern-day legal deposit scheme, and why the domain crawls already conducted by the NLI are inaccessible to researchers and members of the public.[49] Such reasons include the need to have consultation with multiple stakeholders, such as the publishing, heritage communities, and

[48] Copyright and Other Intellectual Property Law Provisions Act 2019, https://www.irishstatutebook.ie/eli/2019/act/19/enacted/en/index.html

[49] Examples here are added in the Bibliography and include: Dáil Éireann. (2021). Digital Archiving; Dáil Éireann. (2021). Intellectual Property; Joint Committee on Tourism, Culture, Arts, Sport and Media. (2021b). Key Priorities and Legislation of the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media; Seanad Éireann. (2021). An tOrd Gnó - Order of Business; Seanad Éireann. (2021). Nithe i dtosach suíonna - Commencement Matters – Digital Archiving.

other government departments. There is also the question of the capacity of the NLI, and thus, a feasibility study should include details of the capacity of the NLI to take on a digital legal deposit web archive, in terms of infrastructure, technology, human resources, etc. For example, on 09 September 2021, in a Dáil Éireann session on parliamentary questions, TD Rose Conway-Walsh (Sinn Féin) put a question to the Minister for Tourism, Culture, Arts, Gaeltacht, Sport and Media, who at the time was (and still is at the time of writing) TD Catherine Martin (Green Party), requesting when the Minister would produce the "report on the feasibility of establishing a digital legal deposit scheme for large-scale, systematic and sustained archiving of the Irish web domain" (Rose Conway-Walsh, Dáil Éireann, Digital Archiving, 09 September 2021). Minister Martin responds as follows:

> My Department is working with the National Library of Ireland (NLI) on exploring the feasibility of expanding the NLI's capacity to establish a digital legal deposit scheme to serve as a web archive for the .ie domain and work is ongoing. There are differing viewpoints on the introduction of digital legal deposit and it is important that consultations incorporate all viewpoints. It is hoped to bring forward a report in the coming months (Catherine Martin, Dáil Éireann, Digital Archiving, 09 September 2021).

Lyman (2002) also points to the need for "a process of negotiation among interested parties" when developing a web archive.

The question of the feasibility report was addressed again in a Seanad Éireann session on 30 Sep 2021, when Senator Fintan Warfield (Sinn Féin) drew attention to the inability of the NLI to archive the web as part of digital legal deposit, remarking that the "loss of digital material means that there is going to be a black hole in our nation's memory" (Fintan Warfield, Seanad Éireann, An tOrd Gnó, 30 Sep 2021). On 11 November 2021, during a Dáil Éireann debate, TD Imelda Munster (Sinn Féin) requested information from Minister Martin on the status of the progress of the feasibility report and reminded the government of their obligations to deliver a feasibility report under section 108 of COIPLPA, 2019 (Imelda Munster, Dáil Éireann, Intellectual Property, 11 November 2021). Minister Martin's response is outlined below

> My Department is working with the National Library of Ireland (NLI) on exploring the feasibility of expanding the NLI's capacity to establish a digital legal deposit scheme to serve as a web archive for the .ie domain and work is ongoing. There are differing viewpoints on the introduction of digital legal deposit and it is important that consultations incorporate all viewpoints. It is hoped to bring forward a report in the coming months (Catherine Martin, Dáil Éireann, Intellectual Property, 11 November 2021).

Several days later, in Seanad Éireann on 23 November 2021, Senator Warfield (Sinn Féin) highlights how "a black hole will be created in our country's memory" due to the failure of putting a legal deposit scheme in place for archiving the .ie domain, and asks "the Minister's view of what the scheme should look like? When will the report be brought to Cabinet?" (Fintan Warfield, Seanad Éireann, Nithe i dtosach suíonna, 23 November 2021).

Warfield further noted how there seems to be "no urgency from Ministers to take ownership of the issue and to set up a digital legal deposit scheme." Standing in for Minister Martin with a written reply, Deputy, TD Peter Burke (Fine Gael), the Minister of State at the Department of Housing, Local Government and Heritage, at the time, delivered what he claims to be a scripted answer from Minister Martin, which first describes the purpose, and remits of the National Library of Ireland regarding this matter, and he then adds the following:

> Legislation could be introduced to give the library the right to conduct a full domain trawl of all .ie websites of Irish interest periodically. To capture a complete record of Irish websites, the domain trawl would include the collection of content behind paywalls. The intention would be that the NLI would make the content available on its premises, as with other resources. This is not a simple issue. However, the owners of websites whose content lies behind a paywall have rights as publishers in general and are important stakeholders in that context. The agreement of relevant publishers would be appropriate and desirable in respect of any legislation (Peter Burke, Seanad Éireann, Nithe i dtosach suíonna, 23 November 2021, my underline).

There are a few points of interest here. First, the script read by the Deputy on behalf of the Minister, uses the term 'domain trawl' instead of 'domain crawl', on two occasions, which implies either a lack of knowledge by the Minister of the fundamental terminology for the issues at hand, or there was a typo. The second point is the Deputy's reference to the need for the agreement of "owners of websites whose content lies behind a paywall have rights as publishers in general and are important stakeholders". With the same logic then, one would also assume that there will be a need for discussion and negotiations with other types of stakeholders. For example, representatives from the teaching and education sector, as well as the end users who use web archives such as academics from a wide range of disciplines, and other types of end users such as public administrators, journalists, legal professionals, web designers, computer scientists, data analysts and local historians (section 4.4.1 & 4.5.1; Ramesh & Hern, 2013; Winters, 2017; Truman, 2016; Bailey, 2015).

The question of the feasibility report was addressed again by Senator Warfield (Sinn Féin) on 24 November 2021, at a Joint Committee on Tourism, Culture, Arts, Sport and Media

meeting. Minister Martin attended this meeting to discuss Key Priorities and Legislation of the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media. Here, Warfield addressed the Minister with another reminder to the government of their obligations to deliver a feasibility report and that they were "breaking the law in not bringing forward that report" (Fintan Warfield, Joint Committee on Tourism, Culture, Arts, Sport and Media, 24 November 2021). The Minister first tried to bypass giving an answer to the question, in which Warfield must repeat the question again. But then, what can only be described as very worrying indeed, is the short response by the Minister as follows:

> The National Library of Ireland is already doing work to digitise websites. We will continue to work with it and the National Archives with a view to addressing the question of digital archives (Catherine Martin, Joint Committee on Tourism, Culture, Arts, Sport and Media, 24 November 2021).

There are several points of interest which can be derived from this response. First, the Minister's response seems to imply that the work that is currently being conducted by the NLI through a low-scale selective web archiving programme (which is currently the only approach allowable by law) can in some way be justified as a reasonable attempt to preserve Irish digital heritage for future generations vis-à-vis the reality of the mass losses of Irish web heritage from the national record.

In September 2022, the NLI Web Archive had 3,105 captured websites (since 2011) in the selective public NLI Web Archive (see Figure 5.7), yet at the end of 2021 there were 330,108 .ie domains registered in the IE Domain Registry (.IE Domain Profile Report, 2021). The statistics from the IE Domain Registry also show an increase of more than 100,000 new .ie domain registrations since 2016 (Figure 5.13).

.com	.net	.org
www.irishtimes.com	www.catholicireland.net	www.one-dublin.org
www.tourismireland.com	www.eircom.net	www.ouririshheritage.org
www.corkairport.com	www.fencingireland.net	www.ireland.anglican.org

Figure 5.13: Screenshot of graphs from the IE Domain Registry, representing the years from 2016 to 2021, for the total number of .IE domain names registered and the number of new registrations for .ie domain names

Furthermore, that does not account for the thousands of Irish websites registered under other domains such as .com, .net or .org. (see examples below). Thus, it is hard to fathom how a Minister who is charged with looking after 'culture', 'arts', and 'media' would consider that the selective web archiving of 3,150 websites since 2011 out of a pool of 330,108 websites registered in the .ie domain 'alone' in 2021 can be justified as a makeshift solution to curb the annual haemorrhaging of Irish digital heritage into the digital dustbin. This also brings into question the actual parameters being decided for what should be included in an Irish national domain web archive, and who gets to decide?

The emphasis through the Oireachtas debates seems to be that a national web domain archive should include a .ie domain crawl, but, as has already been shown, basing a national web archive domain solely on a country's ccTLD, like .ie is too minimal as a representative marker for the collection of a country's national web estate (Webster, 2019; Day, 2003; Coram, 2015). So, there is a need to evaluate the parameters for what should be included in an Irish national web domain archive from the outset and accounted for in the legislation. For example, websites outside the .ie domain (e.g., .com, .org) could be further demarcated if the website contains an ROI postal address, if the website owner confirms residence at an ROI postal address, if the website lists its place of business as an ROI postal address, if the website is identified as being hosted on a server located physically in the ROI via a geo-IP lookup, or if the website focuses on the Irish language, or uses a hybrid approach towards English/Gaeilge. Consideration should also be given to requesting permissions from website owners for websites beyond the realms of the ROI web space which reflect the Irish language or if the website belongs to an Irish immigrant or a diaspora community. Websites which

have a variety of Creative Commons licences might also be considered for inclusion. Furthermore, the users of web archives should also be involved in the discussion of the selection criteria for what gets included in a national web domain archive, as pointed out by Jatowt et al. (2008).

Second, as pointed out in Chapter 3.0, there are several challenges with permissions-based selective collections. Not all websites provide contact details and even if a contact is found there is no guarantee that a website owner will respond (Ryan et al., 2022; Bingham & Byrne, 2021). Pennock (2013) and Brown (2006) point to the weaknesses of selective web archiving due to selector bias (albeit it unintentional or unacknowledged). Third, Brown (2006) notes how the sheer size and depth of the web, makes it difficult for manual selectors to stay abreast of evolving sources, and subject knowledge. Fourth, as pointed out in Chapter 2.0, the very fact that the concepts of in-groups and out-groups are acknowledged as a phenomenon of human behaviour which exhibits in-group favouritism, and discrimination towards out-groups (Tajfel 1970; Tajfel 1971; Tajfel et al., 1974) will also influence what is included or excluded as part of a selective thematic collection. Finally, chapter 2.0 argued that legal deposit libraries across Europe opt to conduct both selective and domain-wide web archiving to combat these issues, achieving a more balanced, representative, and inclusive approach towards the capture of national digital heritage on the web.

From a review of the debates, it appears that the Department of Tourism, Culture, Arts, Sport and Media is responsible for producing a feasibility report that would outline the necessary requirements to inform the Department of Enterprise, Trade and Employment in the drafting of the necessary copyright/legal deposit legislation. We can also surmise that the Department of Enterprise, Trade and Employment and the NAI are identified as stakeholders with "viewpoints" alongside the NLI and the Department of Tourism, Culture, Arts, Sport and Media, and let us not forget the "viewpoints" of the "owners of websites whose content lies behind a paywall [who] have rights as publishers in general and are important" (Peter Burke, Seanad Éireann, Nithe i dtosach suíonna, 23 November 2021). As previously mentioned, one would hope that there are dialogues taking place with other types of stakeholders - representatives from the teaching and education sector, academics across a multitude of disciplines, and other types of end users such as public administrators, journalists, legal professionals, web designers, computer scientists, data analysts and local historians (Healy et al. 2022, p. 26; p. 102; p. 122; Ramesh & Hern, 2013; Winters, 2017; Truman, 2016, pp. 29–30; Bailey, 2015). It would also be beneficial to hold dialogues with information professionals from other national libraries who have experienced the transition from small-scale selective web archiving to large-scale web domain archiving, can advise on

the challenges inclusive of how legislation impacts on implementation and use (Gooding et al., 2019), and thus, minimise the issues from the start. Most importantly, we need to value the opinions of information professionals with experience of working in Irish libraries and information ecosystems, and the Irish archives sectors. One example of this is evident from a sitting of the Joint Committee on Tourism, Culture, Arts, Sport and Media for a discussion on Engagement with Chairperson Designate of the Board of the National Library of Ireland (Joint Committee on Tourism, Culture, Arts, Sport and Media, 13 October 2021).

During the session, Mr. Eoin McVey, the NLI Chairperson (at the time), and Dr Sandra Collins, the NLI Director (at the time), were invited to discuss the challenges and achievements of the NLI (Joint Committee on Tourism, Culture, Arts, Sport and Media, 13 October 2021). As part of this, TD Johnny Mythen (Sinn Féin) asked the NLI representatives:

> What are the legal obstacles in the way of archiving material through digital content? Does this need to be changed as soon as possible? What adequate supports does the library need that are not in place now? (Johnny Mythen, Joint Committee on Tourism, Culture, Arts, Sport and Media, 13 October 2021).

Dr Collins' reply is worthy of reproducing below:

> This is a critical issue for us. We collect one copy of every book published in the State, through copyright legislation legal deposit. We need to acknowledge the importance of content published on websites. Websites are a record of Irish life and we need to be able to make a copy of them and store and preserve them for future use and access. Section 108 of the Copyright and Other Intellectual Property Law Provisions Act 2019 is important. It allows for a report to be brought to Cabinet on the feasibility of a digital web archive. We are working with our parent Department, the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media, to bring that report to Cabinet. It is critical to us (Collins, Joint Committee on Tourism, Culture, Arts, Sport and Media, 13 October 2021).

Dr Collins further expands on these issues claiming that each year, "50% of Irish websites vanish forever or are changed so that they are unrecognisable from what they are now. The records of referendums and general elections are all gone", and notes how the NLI "will not be able to take the risk of collecting it because of the risk and responsibility that puts on the library in terms of having breached copyright legislation" (Collins, Joint Committee on Tourism, Culture, Arts, Sport and Media, 13 October 2021). Finally, Collins suggests that it

> would be useful for the report to go to the Cabinet for consideration and that the report recommend a legislative amendment to copyright legislation, which is the responsibility of the Department of Enterprise, Trade and Employment.

> That, in time, would allow us to capture those websites and our contemporary history before it is gone forever.
>
> Looking across Europe at our peer national libraries, 60% of European national libraries have this legislation in place and are collecting their countries' websites. We do not want to fall behind and lose the data to a black hole forever (Collins, Joint Committee on Tourism, Culture, Arts, Sport and Media, 13 October 2021).

Dr Collins sums up the situation well and offers more contexts regarding the urgent need to amend legal deposit legislation for the inclusion of a national domain web archive.

Another example is noted in the NLI Collection Development Policy 2022-2026, where the NLI clearly states its concern to "collect, as comprehensively as possible, the record of contemporary Ireland". This "record is largely online and highly ephemeral." It concludes that "urgent attention must be given to introducing legislative provision under digital legal deposit for web archiving at scale. Without this, there is a growing and irretrievable gap in the record of Ireland's history, heritage and recorded knowledge" (NLI, Collection Development Policy 2022-2026). In addition, the NLI has pointed out elsewhere that, while selective web archiving is a beneficial activity in the overall picture of the preservation of Irish digital heritage, it is an approach that must be balanced out with a "routine full-domain" crawl of the Irish web space, for it to be in any way representative of the diversity of Irish digital heritage on the web for future generations (Ryan, et al. 2022).

In the meantime, the question of legal deposit and a national domain web archive has somewhat disappeared from the discourse in the Oireachtas with only a couple of mentions in 2022. It was mentioned during a debate of the Joint Committee of the Irish language, the Gaeltacht and the Irish Speaking Community on 15 June 2022. It was mentioned again on 14 September 2022 by Senator Warfield during a Joint Committee on Tourism, Culture, Arts, Sport and Media debate on the Report of the Future of Media Commission. The Future of Media Commission was established in September 2020 as an

> independent body to undertake a comprehensive and far-reaching examination of Ireland's broadcast, print and online media, and to consider how media can remain sustainable and resilient in delivering public service aims over the next decade (The Future of Media Commission, 2022, p. 2).

Warfield's comments are worth noting here.

> I welcome the report's suggestion that any new works should be archived. This committee frequently discusses archiving when representatives from the National Archives come in. The digital legal deposit is not mentioned in this report. Given how much archiving is mentioned in the report, I cannot let the

> opportunity pass to ask where we are with the digital legal deposit that would enable the systematic capturing of the ".ie" domain by the National Library, for example [...] I am blue in the face from raising this over the years as we lose our country's memory as it disappears from the web (Fintan Warfield, Joint Committee on Tourism, Culture, Arts, Sport and Media, 14 September 2022).

Of interest here is that digital legal deposit was not mentioned in the *Report of the Future of Media Commission*, which seems strange, considering the fact that the commission had an undertaking to examine "Ireland's broadcast, print and online media, and to consider how media can remain sustainable and resilient in delivering public service aims over the next decade" (The Future of Media Commission, 2022, p. 2). This is perhaps no surprise. As noted earlier, Lyman (2002) suggests that "in times of innovation the focus is on building new markets and better technologies" rather than solving the "problem of finding a business model to support new media archives" (p. 39).

5.4.3 Web Archives in the Irish media

In a reflective article titled 'Breaking in to the mainstream', Winters (2017) discusses the role that media and newspapers can play in highlighting the value and importance of web archives, and draws attention to one of the first examples of web archives being mentioned in the UK "news rather than technology pages" (p. 175). Here, Winters (2017) refers to an incident which was reported in *The Guardian* newspaper on November 2013, whereby the Conservative party,

> deleted more than a decade's worth of speeches from its website. The story was given an added news angle because one of those speeches was by the then Prime Minister David Cameron praising the Internet for 'making more information available to more people' (Winters, 2017, p. 175).

The article in *The Guardian* also noted that the party also took steps to block access to the captured web pages in the Internet Archive's Wayback Machine (Ramesh & Hern, 2013). Although it is worth noting that the Conservative party website had also been archived by the UK Web Archive, and so the missing web content "had been preserved as a part of the national historical record" (Winters, 2017, p. 175). Nonetheless, for Winters (2017), "the media, and newspapers in particular, have an important role to play" in making "the case for the significance of web archives" (p. 175).

In an Irish context, web archives have been mentioned in the Irish mainstream media on a few occasions, but mainly in specialised sections such as technology. Indeed, it is difficult to find examples where web archives entered the discourse of the mainstream "news" until

recently, when Hugh O'Connell reported in the *Sunday Independent* how Sinn Féin wiped "years of media statements" from their website, but also how the missing media statements were available in the Wayback Machine. In the next section we look at some of the Irish media which has mentioned web archives, and finish with a discussion on the Sinn Féin website incident.

One of the earliest examples of web archives in Irish media is by Michael Cunningham in the 'COMPUTIMES' section of *The Irish Times*, and discusses the work of the Internet Archive for the preservation of the world wide web, and he asks: "If a digital national archive is important for the historians of the future, where is Ireland's digital archive?" (Cunningham, 1997a, p. 18). While not referring to web archives specifically, but born digital data in general, in the Business section of *The Irish Times*, Kieran Fagan (2012) discusses the lack of preservation mechanisms in place for Irish born digital heritage, with a focus on the lack of preservation of the digital records of the Irish government in all their manifestations (Fagan, 2012). In 2017, in the Technology section of *The Irish Times*, Charlie Taylor headlined an article with "Ireland's digital content in danger of disappearing, specialist warns", which discussed the importance of archiving the .ie domain, and how the NLI partnered with the Internet Archive to conduct a crawl of the .ie domain in 2017 (Taylor, 2017b). In 2020, the Librarian and College Archivist of the Library of TCD, Helen Shenton, wrote a Letter to the Editor of *The Irish Times* to discuss the "Digital black hole in our national memory" and the failure of the Irish government to include the web archiving of the Irish national domain as part of the Copyright and Other Intellectual Property Law Provisions Act 2019 (Shenton, 2020). However, to my knowledge, it is not until more recently that web archives entered the Irish mainstream news.

On 13 March 2022, Hugh O'Connell broke a somewhat sensationalist story in the *Sunday Independent* titled: 'Sinn Féin wipes years of media statements from website'. In O'Connell's words:

> Sinn Féin has deleted thousands of media statements that go back nearly two decades from its website in recent days, the Sunday Independent can reveal. The purging of thousands of comments by party representatives, including leader Mary Lou McDonald and her predecessor Gerry Adams from its official website, comes amid controversy over Sinn Féin's previous positions on Russia and its calls for the abolition of Nato (O'Connell, 2022).

The Irish Independent, *The Irish News* and *The Journal* also picked up on the story, with everyone seemingly pointing a finger at Sinn Féin for deleting media statements regarding Sinn Féin's position on Russia and NATO, following the Russian invasion of Ukraine at the

end of February 2022 (Gataveckaite, 2022; Finn, 2022a; *The Irish News*, 2022; O'Connell, 2022).

Picking up the story for *The Irish Independent*, Gataveckaite provided a quote from a Sinn Féin spokesperson claiming that the disappearance of the content is due to "the process of building a new website and archiving outdated content" (spokesperson cited in Gataveckaite, 2022). Finn chased the story for *The Journal* and quoted the Sinn Féin leader Mary Lou McDonald claiming that the deletion of the web content from the Sinn Féin website was "not an attempt to pivot Sinn Féin's position" from issues such as Russia and NATO (McDonald cited in Finn, 2022a). McDonald continued to explain that "'The website is getting a long overdue overhaul. So the archives are being changed'" (McDonald cited in Finn, 2022a). Finn notes how McDonald makes light of the problem claiming "'that there is nothing of note about it'" (McDonald cited in Finn, 2022a). However, McDonald's next response to the incident is quite baffling, as she continued to be quoted by Finn (2022a):

> You don't remove things from the internet, when something is issued, it is there forever, you don't have to be a Pulitzer Prize winning journalist to find comments, remarks or statements on anything. So that's just a housekeeping matter (McDonald cited in Finn, 2022b).

Disappointingly, the leader of Sinn Féin seems to believe that the internet is some type of self-preserving archive that manages to mysteriously keep everything "forever", even though scholars have spent the last quarter of a century drawing attention to the rapidity to which websites disappear from the live web, and unless they are archived, they are lost "forever".

Furthermore, the Sinn Féin leader's statement contradicts the positions of Senator Warfield and other Sinn Féin politicians who have been lobbying for the establishment of a national web domain archive since at least 2018, due to the continual losses of Irish digital heritage from the web, "contrary to the common narrative, what goes online does not stay there forever" (Fintan Warfield, Seanad Éireann, An tOrd Gnó, 30 Sep 2021).[50] Indeed, the Sinn Féin leader could have easily put the story to bed by indicating that the Sinn Féin website had been regularly archived by the NLI since 2011, so there was a 'state' record of those media statements. Figure 5.14 offers an overview of the Sinn Féin website in the NLI Web Archive. In addition, the Sinn Féin website was also archived by the UK Web Archive in 2010, 2014, and every year thereafter, to date. More importantly, McDonald could have used the

[50] An tOrd Gnó - Order of Business – Seanad Éireann Debate – Thursday, 30 Sep 2021, https://www.oireachtas.ie/en/debates/debate/seanad/2021-09-30/8

opportunity to divert attention to the bigger issue here, being the failure of the Irish Government to deliver a feasibility report to establish an Irish national domain web archive, which would ensure that political party websites are archived, regardless of "housekeeping matters" (McDonald cited in Finn, 2022b). This example also demonstrates that amongst many Irish politicians and policy makers there is a lack of awareness regarding the value and need for the preservation of digital heritage on the web. There also appears to be a lack of realisation that the longer Irish politicians and policy makers delay in delivering a report, and amending the legislation, they are ultimately responsible for contributing to a deafening silence in Irish heritage for future generations.

Figure 5.14: Screenshot of NLI Web Archive interface, showing multiple captures of the Sinn Féin website from 2011 to 2022 (www.sinnfein.ie), taken on 2022-10-06

5.4.4 ROI In Brief

As it stands, the ROI is already "impoverished" (UNESCO, 2003) due to mass losses of digital heritage on the web for the decades of the 1990s, 2000s, and 2010s. It now looks like this will continue well into the 2020s, before the necessary measures are put in place for the collection and preservation of the web space of the twenty-six counties of the ROI in line with the collection and preservation of the web space of the six counties of NI. The parliamentary questions and answers above clearly indicate how the loss of Irish digital heritage to "posterity" is not only due to the lack of a business model for the preservation of digital media as digital heritage and supportive legislation, but also due to political circumstances. Moreover, there is no sense of urgency by the government department responsible for delivering a feasibility report to establish an Irish domain web archive. This chapter further suggests that many Irish politicians and policy makers are unaware of the value and need for the preservation of digital heritage on the web and do not seem to realise

how their inactions in establishing a domain web archive are contributing to the problem. It also adds some meaning to why UNESCO (2003) suggests the loss of digital heritage has often gone unnoticed by societies and nations because "Attitudinal change has fallen behind technological change". Consequently, the economic, social, intellectual, historical, and cultural value, or potential value of the heritage is not realised.

It further demonstrates one of the main causes for the loss of Irish digital heritage in the ROI is due to what Lyman (2002) describes as a "cultural problem" whereby "past generations did not, or could not, recognize their historic value" (p. 39). However, surely the Minister responsible for the delivery of a feasibility report on the establishment of an Irish domain web archive realises that the longer they delay in delivering a report, and amending the legislation, they are also ultimately responsible for contributing to the catastrophic loss of Irish digital heritage for current and future generations.

Immediate action is required for an emergency change in the legislation to allow for the collection and preservation of the ROI web estate in the interim, while a feasibility report can continue to be undertaken to advise on the necessary requirements to update the legislation to establish a national web domain archive through "a process of negotiation among interested parties" (Lyman, 2002, p. 40). Moreover, as demonstrated, negotiations should be inclusive of a wide variety of representatives across multiple sectors such as education and teaching, users of web archives for multiple purposes, information professionals with experience in the transition from small-scale selective web archiving to large-scale domain web archiving, and information professionals who are experienced in working with Irish based information ecosystems. Such inclusivity would minimise the challenges from the start.

5.5 Summary

Through a collaborative effort, this chapter engaged with desk research, a review of the literature, and informal dialogues with heritage colleagues to examine the availability and accessibility of web archives based on the island of Ireland, and their usefulness as resources for Irish based research (RQ3). It further examined the causes for the loss of Irish digital heritage (RQ1) and how this contributes to the challenges for participation in web archive research (RQ2); and offered some perspectives on approaches for improving the conditions for conducting web archive research (RQ5).

First, the chapter outlined how several institutions had a responsibility for the collection and preservation of the records and publications of the island of Ireland preceding the digital

turn and described how the destruction of the PROI Record Treasury at the start of the Irish Civil War in 1922 was a cultural catastrophe for the heritage of the island. The chapter further demonstrated how the mass losses of Irish digital heritage from the 1990s to the present day, amounted to yet another cultural catastrophe for the island and emphasised the importance of web archiving as a mechanism for the preservation of national digital heritage.

In dealing with the availability and accessibility of web archive collections based on the island of Ireland, the chapter acknowledged how Irish web heritage can be found in several web archive collections, however we were only interested in web archive initiatives that are based on the island of Ireland and have a specific mandate to capture a wide range of Irish digital heritage as part of their collection strategies. Therefore, the chapter focused on the PRONI Web Archive, the NLI Web Archive, and the UK Web Archive, which is accessible onsite in the Library of TCD. The chapter offered an overview of their historical backgrounds, inclusive of how copyright and legal deposit has influenced their collecting activities and assessed their efforts for the collection and preservation of Irish digital heritage from the web, and their availability and accessibility as resources for conducting Irish based research.

The chapter outlined how there was a balanced approach towards the preservation of the NI web space, through a joint effort by the PRONI web archive, and the UK Web Archive with a wide range of collections covering multiple topics which would be useful for current and future Irish based research. However, it was also noted how there are challenges in the use of the NPLD legal deposit collections in the UK Web Archive due to the restrictive nature of the NPLD access protocols which are outdated in line with advances in publishing and communications technologies, and current trends in digital user expectations and information seeking behaviours (Gooding et al., 2019).

In terms of the ROI web space, the chapter established how the NLI can only operate a small-scale preservational strategy for Irish digital heritage on the web due to the failure of the ROI government to include the web archiving of the Irish national domain as part of legal deposit legislation, and how this failure contributes to mass losses of Irish digital heritage. The chapter discussed the political debates and the inertia for the inclusion of the archiving of the Irish national web domain as part of legal deposit legislation in line with other countries, and how the establishment of a ROI national domain web archive was a necessary component for the preservation of Irish national digital heritage. The chapter emphasised how immediate action is needed to allow for the capture and preservation of ROI web heritage for current and future generations. Once this is secure, negotiations with multiple stakeholders can take place regarding access protocols. In addition, the chapter highlighted

the need to assess the demarcation of the Irish national domain, as using the .ie ccTLD is not an adequate marker for the representation of Irish digital heritage on the web. Moreover, the chapter underscored how born digital content is more fragile than print material, and publishing and communications technologies are constantly changing. Any legislation implemented will therefore have to be reviewed on a regular basis in order to keep up with these changes.

Finally, the chapter highlighted the need for the inclusion of representatives from the teaching and education sectors, academics from a broad range of disciplines, and other types of end users such as public administrators, journalists, legal professionals, web designers, computer scientists, and local historians. It further suggested the need for dialogues with information professionals who have experience in the transitions from small-scale web archiving to large scale domain web archiving, as well as with information professionals who are experienced in working with Irish based information ecosystems. Legislators generally do not have any expertise in managing born digital content and rely more on the legal advice related to print materials. As outlined in Gooding et. al. (2019) this causes conflict in implementation and use. However, as mentioned in chapter 1.0, there is very little known about the users or potential users of web archives in Ireland, therefore this will be further investigated in chapter 6.0.

6.0 AWARENESS OF, AND ENGAGEMENT WITH, WEB ARCHIVES IN IRISH ACADEMIC INSTITUTIONS

If researchers today want to fully understand the present, as well as our past from the mid 1990s onwards, the Web will play a critical role. While there is no common rule for when a topic becomes 'history', the timeframe seems to be shortening as the speed of information dissemination accelerates (Brügger & Milligan, 2019, p. xxviii).

6.1 Introduction

The previous chapter examined the availability and accessibility of web archives based on the island of Ireland, and their usefulness as resources for Irish based research. In doing so it examined the causes for the loss of Irish digital heritage and how this contributes to the challenges for web archive research across Ireland. It further demonstrated the challenges for the preservation of digital heritage due to political and legal conditions and emphasised the need for the inclusion of multiple representatives in the negotiations for the establishment of an Irish national web archive domain. Such negotiations should be inclusive of representatives from the teaching and education sectors, and end users such as academics, public administrators, journalists, legal professionals, and web designers, as well as experienced information professionals.

As mentioned previously, very little is known about the users or potential users of web archives in Ireland. Indeed, publication of Irish based research integrating the use of archived web content is difficult to find with some exceptions being Malone (n.d.), Harjani (2018), Byrne (2019), Greene & Ryan (2019), Healy (2019), Webster (2019), and Greene (2020). Moreover, to date, there appears to be no web archive user studies conducted across Irish academia which examines scholarly engagement, or awareness of the existence of web archives as resources for research. And, as has been observed by the web archivist at the NLI Web Archive, "'It's difficult to get good analytics on web archive users, due to the fact the selective web archive can be accessed remotely'" (Ryan cited in Vlassenroot, 2019, p. 100). In essence, very little is known about those who engage with, or might potentially engage with, web archives as resources for Irish based research. Therefore, it is difficult to assess what types of support and incentives would be most effective for assisting scholars and educators in the use of the archived web for Irish based research. This calls for further investigation.

With the aid of an online survey, this chapter investigates awareness of web archives, and engagement/ non-engagement with web archives by lecturers, researchers, and students in Irish academic institutions.

The main objectives for the chapter are outlined as follows:

- To investigate the awareness of web archives and archived web content as a resource for study/research,
- To generate a better understanding of the users of web archives in Irish academic institutions, and how and why archived web content is used or not used for study/research,
- To explore the challenges and opportunities for using web archives and archived web content as a resource for study/research.

In pursuit of these aims, the survey focuses on the following research questions:

- What is the current level of awareness for the existence of web archives?
- What are the reasons for a lack of engagement with web archives for research?
- What is the likelihood of a non-user using a web archive for research, after becoming aware of its existence?
- Who are the users of web archives?
- How and why are web archives (and web archived content) used for research?
- What is the perceived value of web archives?
- What is the perceived importance of archiving websites based on specific topics?
- What kind of challenges do scholars perceive for the future use of archived web content in their field of research?

The design of the survey considers the objectives of the study and a review of related literature on web archive user studies with attention to research engagement studies conducted by Costea (2018) and Riley and Crookston (2015).

6.2 Related Literature

6.2.1 Web Archive User Studies

There have been several web archive user studies conducted to date. Some studies focus on engagement with web archives by users in general (Jatowt et al., 2008; Costa & Silva, 2010; Moiraghi, 2018), while other studies focus on scholarly awareness of web archives, and scholarly research engagement and non-engagement (Hockx-Yu, 2014; Riley & Crookston,

2015; Costea, 2018). Other studies focus on web archiving practices, and by correlation, they also examine how web archive collections are being used (Bailey et al., 2014; Bailey et al. 2016; Farrell et al., 2018), and the challenges for research engagement (Truman, 2016; Vlassenroot et al. 2019). A review of some of this literature is available in Chapter 3.0.

6.2.2 Use of Web Archives for Irish Based Research

Malone's (n.d.) contribution comes in the form of a web page titled 'Early Irish Web Stuff'. David Malone was a student and system administrator in TCD School of Maths in the early 1990s and was soon introduced to Unix and the set up for TCD's TCP/IP internet system (Malone, 2016). Realising he had lived through "interesting times", Malone attempted "to try and record or make notes of some of what had happened" (Malone, 2016). This resulted in the development of a web page, titled 'Early Irish Web Stuff', which attempts to track down some of the early Irish websites, and some of the developments happening at the time. Malone consults newspapers, examines conversations in the Usenet archive in Google Groups, consults legacy web pages such as Déjà vu news and the 'What's New' directory on the Mosaic Communications Corporation, and utilises the Wayback Machine to track down the URLs of early Irish websites and web pages (Malone, n.d.).

Using archived social media, Harjani's MSC thesis entitled 'Investigating Information Sharing Behaviour on Twitter: The Case of the Irish Referendum' explores the factors that affect the information people share and consume online using Twitter data archived around the Thirty-sixth Amendment of the Constitution of Ireland or more commonly known as the Repeal of the Eighth Amendment (Harjani, 2018). In doing so, Harjani uses the repost count on Tweets archived as a dependent variable while content-level features were selected as independent variables. The data was then modelled using simple regression methods and supplementary network analysis. The main findings were that

> "negative sentiment is a strong driver of reposts. Conversely, posts by news, campaign and politician accounts do not fare well, exhibiting a negative relationship with repost count. A third finding displays the tendency of campaigners to retweet other campaigns of the same vote endorsement. The consequences map out onto the observed polarisation trend in recent years and the rise of fake news. Some of these findings present evidence in support of present literature posing important theoretical and practical questions for policymakers" (Harjani, 2018, p. 2).

Harjani's research highlights the important role that social media can play in understanding key events in Irish society and highlights the need for discussions regarding the preservation

of social media as national digital heritage, and its inclusion in legal deposit schemes (Harjani, 2018).

Byrne (2019) uses a literature review and three case studies to examine approaches for studying women's sport history, with a particular reference to women's football (soccer), which has been relatively understudied in comparison to male sport histories. The paper reviews digital research methods and offers three potential approaches for studying women's football history, using web archives as a source for research, digitised newspaper collections, and oral histories. Of relevance for this study is Byrne's (2019) demonstration of using a web archive using the SHINE interface. SHINE is a prototype of a potential research tool that can be used to access and analyse web archive data. It was developed as part of the Big UK Domain Data for the Arts and Humanities project funded by the UK Arts and Humanities Research Council. The data that underpins this service was acquired by JISC from the Internet Archive and includes all .uk websites in the Internet Archive web collection crawled from 1996 to April 2013, when NPLD came into effect (UK Web Archive, n.d., JISC UK Web Domain Dataset; Byrne, 2019). SHINE allows users to search the dataset and refine their use of keywords, date ranges and Boolean search terms. Byrne (2019) further offers some examples of Boolean search terms for investigating the history of women's soccer in Ireland.

Greene & Ryan (2019) used a subset of data from the selective NLI Web Archive that was captured in 2016 and developed "a manually-curated core set of 299 popular Irish domains, corresponding to over 68 million web pages, stored as 27,400 individual ARC files". First, they extracted the hyperlinks between all HTML pages in the dataset and looked for link pairs (a source and a target URL). Then they converted each link pair into a domain pair and focus on the domain pairs relating to the core set of 299 domains. They observed that there was "a dense core" of the dataset network that consisted of media, governmental, and sporting websites, and came up with suggestions for future work (Greene & Ryan, 2019).

Healy (2019) discusses the decriminalisation of homosexuality in Ireland in line with developments of the world wide web and explores methodologies for finding and recording early internet and web histories of Irish LGBT+ activism. Healy notes how LGBT+ news and activities originating from Irish based email addresses can be found in the Queer Resource Directory as early as 1992, and posts can be traced to Irish based members in the Usenet soc.motss newsgroup from at least 1991. In relation to the web, Healy discusses how the Wayback Machine is useful for examining websites of Irish LGBT+ organisations from the 1990s onwards and demonstrates how such websites underwent "many transformations not merely due to technology, but in terms of discourse and web content, as a result of

changes achieved in the social and legal landscape, and an increasing liberalisation to discuss topics which were previously Taboo" (Healy, 2019). Healy also discusses how the Irish 2015 Marriage Equality Referendum was a climax of a twenty-year campaign, to ensure LGBT+ citizens were afforded equal rights and protections. Nonetheless, multiple campaign websites have since disappeared from the live web. For Healy, this amounts to the loss of Irish social, cultural, political, and constitutional history, and reinforces the need for web archiving as a solution for preventing such losses (Healy, 2019).

Using a case study of a web estate of Christian churches in Northern Ireland (NI), the historian, Peter Webster (2019) examines the nature of the .uk ccTLD as a proxy for the UK web space. Using publicly available documentation such as directories which list individual parish or congregational websites for Roman Catholic churches, Anglican churches, Presbyterian churches, and Baptist churches, Webster compiled a list of relevant website URLs. Then, using archived web data and hyperlink analysis, Webster (2019) examines the link relationships between NI churches, including "the regional, national and cross-border relationships that they imply" (p. 111). For this, Webster (2019) uses a dataset made available in the UK Host Link Graph, which is derived from a larger dataset, being the JISC UK Web Domain Dataset and is made available by the British Library.[51] In doing so, Webster (2019a) draws attention to the difficulties in delimiting the UK web domain solely using the .uk ccTLD as a proxy, due to the vast amount of website content which exists outside of those parameters. For example, UK websites hosted on .com, or .org (Webster, 2019, p. 112). Webster finds that out of 100 domains for Roman Catholic churches in NI, only 12 were registered with a .uk domain, while 3 were registered with the Republic of Ireland's .ie domain. Webster (2019) finds that the links relationships show a very loose mapping to the UK ccTLD (p. 111), and thus Webster (2019) suggests that "for web archivists and scholars alike the ccTLD is a weak proxy indeed for the national web" (p. 120).

Greene (2020) offers a useful demonstration of network analysis using WARC data from the 2018 Irish Presidential Election captured in the public NLI Web Archive. Consisting of 1,000 WARC files, containing 57,065 HTML pages, Greene (2020) extracted links from each page, and mapped each link to a pair of domains, with a focus "on pairs of domains for which both the source and target are distinct", thus excluding internal links. Greene (2020) suggests that

> By representing large collections of web pages as a link network, researchers can apply existing methodologies from the field of network analysis. For web

[51] Host Link Graph JISC UK Web Domain Dataset (1996-2010), https://data.webarchive.org.uk/opendata/ukwa.ds.2/host-linkage/

> archives, we can use these methods to explore their content, potentially identifying meaningful historical trends (Greene, 2020, EWA Book of Abstracts)

Greene highlights how the use of network analysis can benefit the collection development of the NLI selective web archive, as well as for studying the "archived Irish web" (Greene, 2020, EWA Book of Abstracts, 2020).

The literature above provides a useful starting point when considering the type of research that has already been undertaken using web archives for research on Irish based topics, and how it can be built upon. It further demonstrates the use of a qualitative approach (Malone, n.d., Healy, 2019), a big data approach (Greene & Ryan, 2019; Greene, 2020) and combining qualitative and big data approaches (Harjani, 2018; Byrne, 2019; Webster, 2019). This provides a good indicator on the types of research which need to be accounted for in any forthcoming legislation on copyright and legal deposit. Moreover, Harjani's (2018) research highlights the important role that social media can play in understanding key events in Irish society. Thus, any new legal deposit legislation introduced in the ROI should consider making provisions for the inclusion of social media content.

6.3 Methodology

In this section, the methodological approach for the chapter is laid out, including the design for the online survey, the recruitment process, and the approaches for data collection and analysis. Online (questionnaire) surveys are a research method for gathering information about behaviours, attitudes, values, and experiences across a broad range of research disciplines and can be used as a standalone method or as part of a combined approach (Dawson, 2020, p. 288). As with all research methods, there are advantages and disadvantages which need to be considered for conducting an online survey as a research method for a user study (Wright, 2005; Steber, 2016). The research for this chapter was conducted in compliance with best practice guidelines for the collection and management of research data, as outlined in Maynooth University Research Ethics Policy (2016) and Maynooth University Research Integrity Policy (2016). To note here these policies were updated in 2019 (Ethics) and 2021 (Integrity), after the data had been collected and analysed. However, this did not affect the research plan. The collected/analysed data will be migrated to a private server repository in Maynooth University, for long-term preservation, for a period of ten years, after which it will be deleted in full (as outlined in MU Research Integrity Policy, 2021).

6.3.1 Survey Software

The data was collected using the SurveyMonkey online survey tool. At the time the research was being carried out, Maynooth University did not have a specific policy for conducting online surveys, or a policy on which type of software to use for such studies. Thus, the researcher opted to use SurveyMonkey due to having prior experience in using the software and having an account subscription. To note here, Maynooth University only introduced a policy for online surveys in November 2019, which specified JISC Online Surveys as the only tool permitted by the university for conducting studies of this nature.

6.3.2 Survey Recruitment

The survey was accessible via a web link inserted in recruitment emails and several posts on the researcher's social media platforms (Facebook, Twitter, and LinkedIn). From November to December 2018, 970 recruitment emails were sent to academics (mostly head of departments, or head of degree programmes) and to department administrators in all fields of research at nine universities in the Republic of Ireland (ROI) and Northern Ireland (NI). The survey was open from November 2018 until January 2019. Table 6.1 provides the list of universities, and a breakdown of the number of emails sent per university.

As well as providing the survey link, emails provided information on the purpose of the study, with an assurance for anonymity and confidentiality. See Appendix D for an example of the recruitment email. As a degree of self-selection bias was expected due to the interests of those who are aware of, or use, web archives, the recruitment email also emphasised the equal importance of participation from respondents who were not aware of or did not engage with web archives.

Early in the email recruitment process, it was noticed that a few of the complete surveys had inconsistent responses with regard to the awareness of and use of a web archive. For example, when participants were asked to name any other web archives that they were aware of or engaged with, some respondents provided the names of digital archives or digital libraries. At this point it was decided to provide some additional text in the recruitment email, briefly noting the difference between a web archive, and a digital archive/library.[52] However, there were similar instances of inconsistencies in later survey responses. This will be discussed in more detail later on.

[52] Additional Note: A 'web archive' is a resource that captures and preserves websites, blogs, and web pages, and provides access to view such content, long after it has disappeared from the

Table 6.1: Breakdown of recruitment emails sent per university

Third-Level Academic Institutions	Sent to Academics	Sent to Admins	Total per Uni
Dublin City University	=49	=37	=86
Maynooth University	=54	=42	=96
National University of Ireland, Galway	=77	=37	=114
Queen's University Belfast	=79	=26	=105
Trinity College Dublin	=102	=40	=142
University College Cork	=70	=45	=115
University College Dublin	=116	=33	=149
University of Limerick	=49	=36	=85
Ulster University	=55	=23	=78
Totals	(=651)	(=319)	(=970)

6.3.3 Survey Design & Questions

The design of the research and survey questions considered the aims of the chapter and a review of similar web archive user studies (section 3.2). An effort was made to ensure the survey was answerable in 8-10 minutes, to increase the chances of completion (Chudoba, 2018; CoolTool, 2017; Steber, 2016). The survey was field-tested by four academic colleagues to ensure the questions were clearly understood, after which some amendments were introduced to the survey language and layout. A final draft of the research project including information about the project, informed consent, how the data would be collected, managed, and used (see Appendix E) and a copy of the survey questions (see Appendix F) was submitted to Maynooth University research ethics committee for approval. The study received ethics approval [SRESC-2018-083] in October 2018.

The survey consisted of thirty questions, but respondents were not required to answer every question. This was dependent on whether a respondent was a user or non-user of web archives for their research/studies. The questions contained a mix of dichotomous, trichotomous, and multiple-choice questions (some with options for free text), Likert scales,

live web. A web archive differs from a digital archive/ library in so far as a web archive only contains archived websites, blogs, and web pages.

and an optional open-ended question at the end. All participants were asked to answer some demographic questions based on nationality, age, gender, position, and discipline category. All participants were provided with a description of an online public and dark web archive and were asked to answer questions on awareness of their existence, and whether they used web archives for their research or studies. Based on a Yes or No answer, participants were then directed to a set of questions for users or non-users. Non-user questions focused on reasons for non-engagement, and the likelihood of engagement with web archives in the future. Users were asked questions on their reasons for using web archives and the web archive resources they engaged with. In the final section, all participants were asked to answer questions on their perceived value of web archives, the importance of archiving websites based on different topics, and the significance of web archives as resources for current, medium, and future use in their field of research. The survey ended with an optional open-ended question to allow participants to comment on their perceptions of the challenges for using web archived content in their disciplines in the future.

It is worth discussing here some of the terminology choices that were made for the types of web archive collections in archiving initiatives. The term "online public web archive" is used to describe a resource "whereby access is available to the general public via the web/internet from any location"; and the term "dark web archive" refers to a resource with no public access or with restricted access "onsite in a designated reading room or Library via an onsite portal." The term dark (domain) web archive is used to refer to an archived web domain collection which has no public access or has restrictive access. For example, regarding the NLI, we use the term online public NLI web archive (selective web archive collection) and the NLI dark (domain) web archive (a collection of domain crawls conducted by the NLI). To note, the domain crawls were conducted in 2007 and 2017, and were at the time speculated to become accessible onsite in the NLI reading room (Taylor, 2017a), but were inaccessible at the time the study was conducted.

The reasons for choosing these terms were first guided by the need to come up with terms that would describe the status of web archives with restrictive access to an unfamiliar audience. We were also guided in some way by the use of the term domain dark archive in the Analytical Access to the Domain Dark Archive (AADDA) project. For example, when describing the AADDA project, Webster (2012) describes it as an 18-month project in the UK which sought to "enhance the sustainability of a substantial dark archive of UK domain websites collected between 1996 and 2010 by the Internet Archive, copies of which were recently acquired by the JISC and are stored at the British Library on their behalf" (Webster, 2012; Analytical Access to the Domain Dark Archive, 2012+). On discussing the use of this

data, Gorsky (2015) describes the dataset as “a large number of UK domain websites, captured 1996–2010, which is colloquially termed the Dark Domain Archive while technical issues surrounding user access are resolved” (p. 596). Thus, it was considered that archived web collections which are not publicly accessible, or which have restrictive access in place may be referred to as a ‘dark web archive’, and archived web domain collections which are not publicly accessible or have restrictions in place may be termed as a ‘dark domain web archive’. From there, it was formulated that the term could be applied to both legal deposit and non-legal deposit archived web domain collections with restrictive access. At the time, it was felt that the use of these terms would be a reasonable way of describing the context of such collections to an unfamiliar audience.

This terminology has, however, evolved since the survey to differentiate dark archives (custodian access only) from dim archives (mix of dark and open), and open archives (light) (Skinner & Schultz, 2010, pp. 128–131; Erickson, 2013), although there are examples where these boundaries are blurred. For example, Lavoie and Dempsey (2004) assert that the “notion of ‘dark archives’, supporting little or no access to archived materials, has met with scant enthusiasm in the library community”, and suggests that dark archives “will function not just as guarantors of the long-term viability of materials in their custody” but also offer “access gateways.” In addition, Martzahl (2010) describes a dark archive as “a secret place for storing archival material with restricted user access.”

6.3.4 Survey Responses

The survey was open from November 2018 until January 2019. 378 participants responded to the survey through email (=367) and social media (=11). However, 93 participants exited the survey prior to completion. This amounted to a completion rate of 75.40%. As participants were informed that their responses would not be recorded if they did not complete the survey, the 93 incomplete surveys were removed and deleted. A further 46 complete surveys were also removed from the survey dataset, due to response inconsistencies, and will be discussed next.

As mentioned previously, early in the recruitment stage, it was noticed that some surveys had inconsistent responses with regards the awareness of and use of a web archive, in so far as some respondents confused a web archive with other types of resources such as digital libraries, digital archives, and data repositories such as Project Gutenberg, JStor and Talkbank. In total, there were 28 such instances. As this study was aimed to address the reasons for user and non-user engagement, it was decided not to include the 28 survey responses in the final tally for analysis. This is also comparable to an occurrence in the study

conducted by Costea (2018). As Costea's study was specifically aimed at users of archived websites, Costea did not factor inconsistent responses into the results. Moreover, there were further instances of inconsistencies in this survey. 18 respondents identified as a user of web archives, yet they indicated that they were not aware of and did not use any of the web archives that were listed, nor did they provide a name of any other web archive they were aware of or used. Riley and Crookston (2015) also came across a similar occurrence in their study of academic institutions; however, they opted to include this data for final analysis, but used filters to calculate their results around the inconsistencies. For this study, however, it was decided not to factor in the 18 surveys with such inconsistencies, so as to provide a clearer representation for users and non-users and the reasons for engagement or non-engagement with web archives. Therefore, the final tally of complete surveys for analysis in this chapter is (N=239).

6.3.5 Survey Limitations

Participation was voluntary, and participants could withdraw at any time during the process of filling out the questionnaire, with the knowledge that their responses would not be recorded. It is also worth noting that some fields of research are more dominantly represented in some universities than others, and some fields of research are not equally available across all universities. Consequently, this may have resulted in an over-representation of participants from some fields of research. It is not possible to evaluate this effect due to ethics considerations, as there were no identifiers collected to evaluate a response rate per university/department. Nonetheless, in the final tally of survey responses for analysis, respondents identified with twenty-four discipline categories, providing a varied range of representations from different fields of research. Also, as with all studies based on survey sampling, this survey study cannot be construed to represent the academic population in Ireland as a whole.

6.4 Results & Analysis

The survey results and analysis are based 239 survey respondents (N=239), and percentages in the discussion and graphs are reflective of this, unless otherwise stated in the case of user and non-user questions and answers.

6.4.1 Demographics

This section provides an overview of responses to questions on nationality, age, gender, position, and discipline categories with some data breakdowns for representations of users

and non-users. The purpose of these questions was to establish some demographic information about participants which might reveal some trends when cross-tabulated with data from other responses in the survey.

6.4.1.1 Nationality, age, gender

Respondents (N=239) identified with 20 nationalities. Table 6.2 offers a representation of participant responses for nationality (N=239), with a comparison of nationality representations for users and non-users. As expected, the highest rate for identification with a nationality was Ireland (75.73%, n=181); of which 44 respondents identified as a user and 137 as a non-user.

Table 6.3 provides an overview of participant responses for age, with a comparison of age representations for users and non-users. Of overall participation (N=239), the highest representation for age is the age bracket of 45-54 (24.27%, n=58) with 14 respondents identifying as a user. This is followed by the age brackets of 18-24 (21.76%, n=52), and 35-44 (20.50%, n=49). Out of the overall participation (N=239), there were slightly more female (52.30%, n=125) respondents than male respondents (45.61%, n=109). Table 6.4 provides a comparison of user and non-user gender representations.

Table 6.2: Representation of participant responses for nationality (N=239), with a comparison of nationality representations for users and non-users

Nationality Answer Choices	Responses (N=239)		User (n=59)	Non-User (n=180)
AT - Austria	0.42%	(n=1)		=1
AU - Australia	0.42%	(n=1)		=1
BG - Bulgaria	0.84%	(n=2)		=2
CA - Canada	0.42%	(n=1)		=1
DE - Germany	3.35%	(n=8)	=5	=3
ES - Spain	1.67%	(n =4)	=1	=3
FR - France	0.84%	(=2)		=2
GB - United Kingdom	5.02%	(=12)	=3	=9
IE - Ireland	75.73%	(=181)	=44	=137
IN - India	1.67%	(n =4)		=4
IT - Italy	2.51%	(n =6)		=6
MW - Malawi	0.42%	(n =1)	=1	
NG - Nigeria	0.42%	(n =1)		=1

NL - Netherlands	0.42%	(n =1)		=1
RO - Romania	0.42%	(n =1)		=1
RS - Serbia	0.42%	(n =1)		=1
RU - Russia	0.42%	(n =1)	=1	
SI - Slovenia	0.84%	(n =2)		=2
UA - Ukraine	0.42%	(n =1)	=1	
US - United States	3.35%	(n =8)	=3	=5

Table 6.3: Representation of participant responses for age (N=239), with a comparison of age representations for users and non-users

Age Bracket Answer Choices	Responses (N=239)		User (n=59)	Non-User (n=180)
18-24	21.76%	(n=52)	=11	=41
25-34	19.67%	(n=47)	=10	=37
35-44	20.50%	(n=49)	=12	=37
45-54	24.27%	(n=58)	=14	=44
55-64	11.72%	(n=28)	=11	=17
65+	1.26%	(n=3)	=1	=2
Prefer not to say	0.84%	(n=2)	=0	=2

Table 6.4: Representation of participant responses for gender (N=239), with a comparison of gender representations for users and non-users

Gender Answer Choices	Responses (N=239)		User (n=59)	Non-User (n=180)
Male	45.61%	(n=109)	=34	=75
Female	52.30%	(n=125)	=24	=101
Other	1.26%	(n=3)	=0	=3
Prefer not to say	0.84%	(n=2)	=1	=1

6.4.1.2 Positions & Disciplines

To better understand the positions of participants within academic institutions, they were provided with a choice of seven position categories or the option of 'Other' to enter free text. 16 respondents provided free text which was coded, of which 10 responses were incorporated into the existing categories that were offered.[53] The six remaining responses were coded into three new categories which are marked with an asterisk (*).These adjustments are included in the final calculations.

Table 6.5 provides an overview of participant responses for position, in line with representations for users and non-users. Figure 6.1 provides a representational graph of the position of users in line with total responses (N=239) and shows that users of web archives within this study are educators (12.55%, =30), researchers (7.95%, =19), and students (4.18%, =10).[54]

In configuring a research/discipline area, participants were offered a choice of 20 categories representing a discipline or collective of disciplines, and the option of 'Other' to enter free text. 29 respondents chose 'Other' and entered free text of which was coded and added to existing categories or incorporated into seven new categories which are marked with an asterisk (*). These adjustments are included in the final calculations.

Respondents identified with 24 discipline categories. Table 6.6 offers a breakdown of responses for the discipline category in line with user and non-user representations. User respondents represent 17 different discipline categories. The number of users vis-à-vis the rate of participation per discipline category shows a strong number of users from the Humanities (21 of 50) but a low number of users from Social Sciences (5 of 33), Engineering Science (3 of 24), and Natural Sciences (2 of 29).

[53] 6 respondents identified as a lecturer to some degree, such as 'Part time lecturer' or 'Lecturer above the bar' - these were added to the existing category for Senior Lecturer or Associate Lecturer; 1 respondent identified as a 'PhD candidate and teaching fellow' – this was added to the existing category of PhD candidate/student; 3 respondents identified as 'Assistant Professor' – these were added to the existing category for Professor or Associate Professor.

[54] Figure 4.1 - Representational Graph: Educator (Senior Lecturer/Associate Lecturer + Professor/ Associate Professor); Student (Undergraduate + Postgraduate + PhD candidate/student); Researcher (Postdoctoral associate, researcher, or fellow + Employed researcher in a third-level educational setting or project).

Table 6.5: Representation of participant responses for position (N=239), with a comparison of position representations for users and non-users

Position Answer Choices	Responses (N=239)		User (n=59)	Non-User (n=180)
Undergraduate student	20.08%	(n=48)	=8	=40
Postgraduate student	10.46%	(n =25)	=4	=21
PhD candidate/student	15.90%	(n =38)	=7	=31
Postdoctoral associate, researcher or fellow	7.53%	(n =18)	=4	=14
Employed researcher in a third-level educational setting or project	4.18%	(n =10)	=6	=4
Senior Lecturer or Associate Lecturer	21.34%	(n=51)	=16	=35
Professor or Associate/Assistant Professor	17.99%	(n =43)	=14	=29
*Administrator (academics/research)	1.26%	(n =3)	=0	=3
*Technical/Support Staff	0.84%	(n =2)	=0	=2
*Director of research centre	0.42%	(n =1)	=0	=1

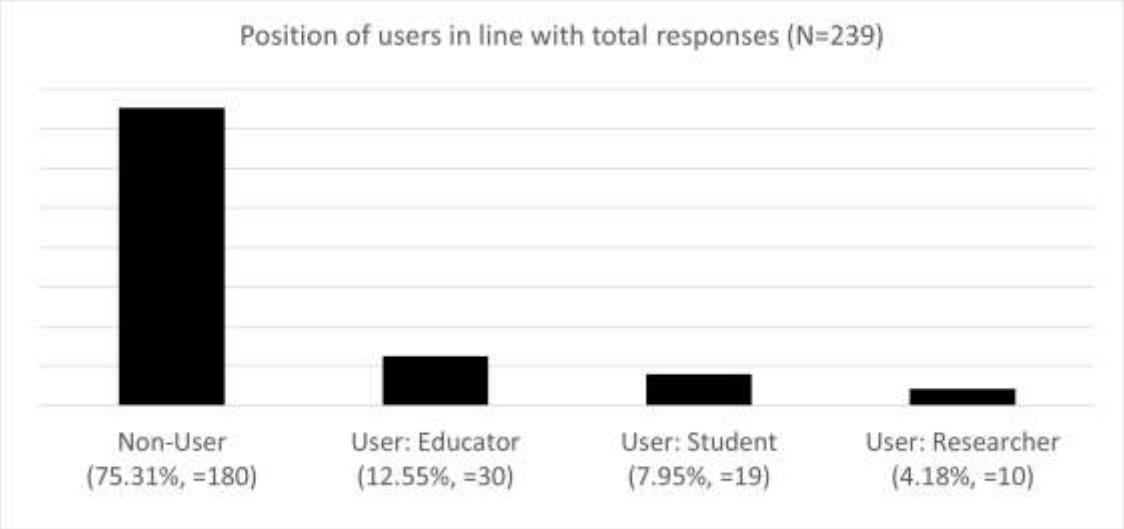

Figure 6.1: Position of users (n=59) under the representations of educators, researchers, and students, in line with total responses (N=239)

Table 6.6: Representation of participant responses for discipline category (N=239), with a comparison of discipline representations for users and non-users

Discipline Answer Choices	Responses (N=239)		User (n=59)	Non-User (n=180)
Architecture	0.84%	(n=2)	=1	=1
Arts (visual, performance, music)	1.26%	(n=3)	=0	=3
Business/Economics/Finance	2.51%	(n=6)	=3	=3
*Built Environment	0.42%	(n=1)	=0	=1
Computer Science	4.60%	(n=11)	=3	=8
*Construction Management	0.42%	(n=1)	=0	=1
*Dental Science	0.42%	(n=1)	=1	=0
Digital Arts/Humanities/Cultural Heritage	1.67%	(n=4)	=1	=3
Educational Science	5.44%	(n=13)	=3	=10
Engineering Science	10.04%	(n=24)	=3	=21
Geography (cartography, hydrology, meteorology, environment)	1.67%	(n=4)	=2	=2
Government/Public Administration	0.42%	(n=1)	=1	=0
*Health Studies/Sciences	4.60%	(n=11)	=0	=11
Heritage Studies, Archival Studies	0.42%	(n=1)	=0	=1
Humanities (history, archaeology, languages, literature, philosophy, theology)	20.92%	(n=50)	=21	=29
Internet Studies	0.00%	(n=0)	=0	=0
Law (criminal, civil, common, statute)	7.53%	(n=18)	=6	=12
Library and Information Sciences	0.00%	(n=0)	=0	=0
Mathematics	1.67%	(n=4)	=1	=3
*Medicine/Biomedical Engineering	0.84%	(n=2)	=0	=2
Media/Communications	1.67%	(n=4)	=3	=1
Natural Sciences (biology, chemistry, physics, earth sciences, space sciences)	12.13%	(n=29)	=2	=27
*Nursing/Midwifery	3.77%	(n=9)	=1	=8
Political Science	2.51%	(n=6)	=2	=4
*Psychotherapy	0.42%	(n=1)	=0	=1
Social Sciences (anthropology, human geography, linguistics, sociology, psychology)	13.81%	(n=33)	=5	=28
Sport and Leisure	0.00%	(n=0)	=0	=0

6.4.2 Engagement with online digital-based resources

Having been provided with four options of online digital-based resources (World Wide Web, Digital Archives, Digital Libraries, and Virtual Research Environments), participants were asked to indicate the frequency at which they use or access these resources for their research or studies. Responses differed for each resource type, but there was a clear indication that 88.28% (n=211) of respondents 'Always', and 11.30% (n=27) 'Sometimes' use the World Wide Web, suggesting that the web is a major research resource in Irish third-level academic institutions. Table 6.7 provides a breakdown for each resource.

Table 6.7: Representation of participant responses for engagement with other online/digital resources

FREQUENCY	World Wide Web (N=239)	Digital Archives (N=239)	Digital Libraries (N=239)	Virtual Research Environments (N=239)
Always	88.28% (=211)	25.10% (=60)	39.33% (=94)	3.77% (=9)
Sometimes	11.30% (=27)	27.20% (=65)	35.15% (=84)	14.64% (=35)
Rarely	0.42% (=1)	20.92% (=50)	12.97% (=31)	28.45% (=68)
Never	0.00% (=0)	26.78% (=64)	12.55% (=30)	53.14% (=127)

6.4.3 Awareness of the existence of web archives

This section provides an overview of responses for awareness of the NLI web archive (online public web archive and dark (.ie) web archive), awareness of other public web archives, and awareness of any other web archives not mentioned. There is also some cross-tabulation of the data with position and discipline representations.

6.4.3.1 Awareness of the NLI Web Archive

With the option of 'Yes' or 'No', participants were asked if they were aware that the NLI archives websites which are made accessible through the online public NLI Web Archive (https://archive-it.org/home/nli). In addition, they were also asked if they were aware that the NLI archived the Irish domain (.ie) in 2007 and 2017 and would soon make it available as a dark web archive – only accessible onsite in a designated reading room at the NLI (Taylor, 2017a; also see NLI, n.d., Irish Domain Web Archive). To note here, at the time this survey was conducted, there was a belief that the archived web domain collections would soon become accessible in the NLI reading room, as noted by Charlie Taylor in *The Irish Times*

(Taylor, 2017a). However, as was demonstrated in chapter 4.0, these collections remain inaccessible due to legalities.

Of all respondents (N=239), 18.41% (n=44) indicated that they were aware of the online public NLI Web Archive and identified their nationalities as Ireland (=33), United States (=4), Germany (=2), and United Kingdom (=1). The 44 respondents identified with 11 different discipline categories, of which Humanities (=21) was the most represented. However, there was a low level of awareness of the resource apropos the rate of participation from respondents from the Social Sciences, Law, Natural Sciences, and Engineering Science (see Table 6.8 for a breakdown). For example, of total participation, 33 respondents identified with the Social Sciences, but only 4 were aware of the online public NLI Web Archive.

Table 6.8: Representation of a comparison of discipline categories of respondents who indicated awareness of the online public NLI Web Archive

Discipline category of respondents	Number of respondents who indicated awareness of the public NLI Web Archive (n=44)	Total number of user/ non-user respondents who identified with that discipline category (N=239)
Business, Economics, Finance	=1	=6
Computer Science	=2	=11
Digital Arts/Humanities/Heritage	=3	=4
Educational Science	=2	=13
Engineering Science	=2	=24
Geography	=2	=4
Humanities	=21	=50
Law	=4	=18
Media/Communications	=2	=4
Natural Sciences	=1	=29
Social Sciences	=4	=33

In the case of the NLI dark (.ie) web archive, 2% (4 of 239) of respondents indicated that they were aware of its existence. The 4 respondents identified their nationality as Ireland; their discipline categories as Humanities (=3), and Law (=1); and their positions as Lecturer (=2), PhD candidate (=1), and Postdoctoral associate/ fellow (=1).

6.4.3.2 Awareness of other online public web archives

Regarding awareness of other online public web archives, participants (N=239) were provided with a list of six international online public web archives (with the URL link to each resource), and options for a 'Yes' or 'No' answer. Table 6.9 provides a breakdown of responses for each resource and shows a highest degree of awareness for the Internet Archive, Wayback Machine (31.38%, n=75); followed by the US Library of Congress Web Archive (23.85%, n=57); the UK Web Archive (21.76%, n=52); the UK Government Web Archive (15.06%, n=36); the PRONI Web Archive (13.81%, n=33); and the UK Parliament Web Archive (12.97%, n=31).

Table 6.9: Representation of participant responses (N=239) for awareness of other online public web archives

Other online public web archives	Responses (N=239)	
	Yes: I was aware	No: I was not aware
Internet Archive, Wayback Machine	31.38% (n=75)	68.62% (n=164)
PRONI Web Archive	13.81% (n=33)	86.19% (n=206)
UK Web Archive	21.76% (n=52)	78.24% (n=187)
UK Government Web Archive	15.06% (n=36)	84.94% (n=203)
UK Parliament Web Archive	12.97% (n=31)	87.03% (n=208)
US Library of Congress Web Archive	23.85% (n=57)	76.15% (n=182)

Participants were also asked if there were any other web archives (not listed above) that they were aware of and offered an option to enter free text. 11 respondents provided free text, and their comments are summarised below.

- While acknowledging that it was not the same as a web archive, 1 respondent mentioned the revision histories in Wikipedia: "Not exactly an archive, but Wikipedia does preserve accessible records of page revisions with data concerning who edited pages and why. This has been important for my research as people sometimes use this as a way to put information into the public domain that the public might not otherwise know to query" (User, Dental Science).

- 2 respondents suggested Google Cache, as a means to retrieve an older version of a website.
- 2 respondents mentioned electoral/referendum collections in the NLI Web Archive
- 1 respondent suggested the Austrian National Library, Web Archive.
- 1 respondent referred to the web archive of the Bibliothèque nationale de France.
- 1 respondent noted Archive Team and Zone-H.
- 1 respondent mentioned Archive.today.
- 1 respondent referred to the List of Web archiving initiatives, Wikipedia page.
- 1 respondent added: 'If this helps...it is not a web archive, but a digital archive I am using: https://curia.europa.eu/en/content/juris/c2_juris.htm' (Non-user, Law).

6.4.4 Engagement with web archives for personal and research interests

In order to initiate inquiry into the reasons for the use of, or non-use of web archives for research, all participants (N=239) were asked: (i) if they ever accessed or used an online public web archive for their personal interest; and (ii) if they ever accessed or used an online public web archive, or dark web archive to assist with their studies or research.

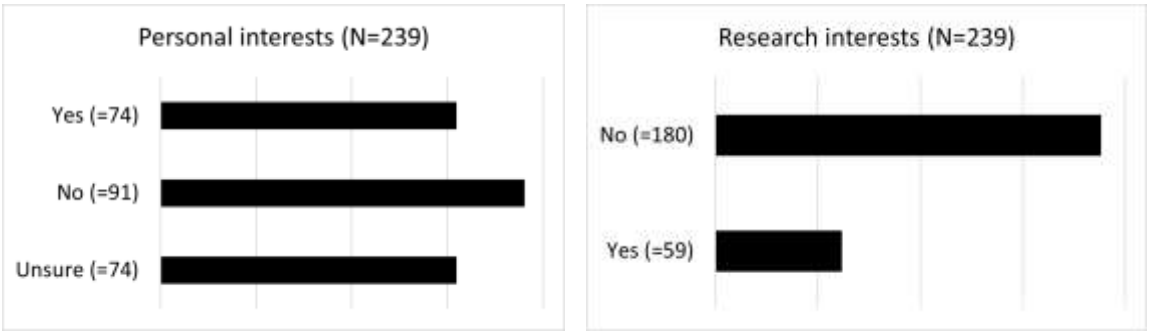

Figure 6.2: Representation of participant engagement with web archives for personal interests and research

As regards the use of an online web archive for personal interests (N=239): 30.96% (n=74) indicated 'Yes'; 38.08% (n=91) indicated 'No'; and 30.96% (=74) indicated 'Unsure' (Figure 5.2). For the use of web archives for their studies or research (N=239): 24.69% (=59) indicated 'Yes', and 75.31% (n=180) indicated 'No' (see Figure 5.2). Respondents were then directed to the corresponding sections for users and non-users.

6.4.5 Non-users of Web Archives for Research

This section provides an overview of responses by respondents who identified as a non-user (n=180) on questions related to the reasons for a lack of engagement with web archives, and the likelihood of future engagement with online public web archives and a dark web archive.

6.4.5.1 Reasons for lack of engagement with web archives

Respondents (n=180) who indicated that they did not access/use web archives for their research/studies were first asked about their reasons for not using an online public web archive for their studies or research. Participants were provided with seven answer choices, an option of 'Other' to enter free text and were asked to tick all that applied. Table 6.10 provides a breakdown of non-user responses (n=180) and shows that a large majority (78.33%, =141) do not engage with web archives for their research, due to a lack of awareness of the existence of web archives.

Other reasons include a lack of knowledge in how to use a web archive (41.67%, =75); how to find archived websites in a web archive that are relevant to a research area (45.00%, =81); uncertainty of the credibility or authority of using archived websites as a primary source (26.11%, =47); and how to cite/reference an archived website from a web archive (17.78%, =32).

26 respondents ticked 'Other reason(s)' for not using an online web archive, and 24 respondents provided free text responses and are summarised as follows,

- 8 respondents indicated that they were unsure as to how relevant, useful, or beneficial, a web archive would be for their research.
- 16 respondents indicated that a web archive was not relevant for their research, of which 4 noted their research required up-to-date sources, and 3 identified as early to modern period historians that required alternate archival sources.

Table 6.10: Representation of non-user respondent (n=180) reasons for not using an online web archive for their studies/research

Answer Choices	Non-User Responses (n=180)
I was not aware of the availability of web archives as resources for my studies/research	78.33% (=141)
I do not know how to use a web archive for my studies/research	41.67% (=75)
I feel that I do not have the technical skills to use a web archive for my studies/research	7.78% (=14)
I do not know how to find archived websites relevant to my studies/research in a web archive	45.00% (=81)
I do not know how to cite/reference an archived website from a web archive to include in my studies/research	17.78% (=32)
I am unsure of the credibility or authority of using archived websites as a primary source for my studies/research	26.11% (=47)
I am unsure about copyright implications for using archived web content for my studies/research	13.33% (=24)
Other reason(s) for not using an online web archive for your studies/ research (please specify)	14.44% (=26)

There is no denying the fact that web archives are simply not relevant for some fields of research. On the other hand, the findings suggest that the value of web archives as a research resource is not clearly understood by an unfamiliar audience. For example, for some respondents, there is a need for more efforts to demonstrate the importance of archiving the web and to promote the value of web archives for research. It could be further suggested that there is a need for the dissemination of use cases in Irish based research that would demonstrate theoretical and methodological approaches for using web archives as a research resource. Of additional interest are the discipline categories of the non-user respondents (=141) who identified with a lack of engagement with web archives, due to a lack of awareness. As mentioned previously, it was surprising to find a low-level of NLI web archive users who identified with the Social Sciences (4 of 33). However, a high-level of respondents from the Social Sciences (22 of 33) and Natural Sciences (20 of 29) do not engage with web archives for research due to a lack of awareness of their existence. A similar case could be made for respondents who identified with some other discipline categories. Table 6.11 offers a breakdown of this data vis-à-vis discipline category.

Table 6.11: Representation of discipline categories for non-user respondents who indicated a lack of research engagement with online web archives due to a lack of awareness (n=141), in line with the total number of user and non-user participants who identified with that discipline category

Discipline categories of non-user respondents, who indicated a lack of research engagement with public web archives, due to a lack of awareness	Number of respondents who indicated non-engagement, due to a lack of awareness (n=141)	Total number of user and non-user respondents who identified with that discipline category (N=239)
Architecture	=1	n=2
Arts	=3	n=3
Business, Economics, Finance	=3	n=6
Computer Science	=7	n=11
Digital Arts/Humanities/Heritage	=2	n=4
Educational Science	=10	n=13
Engineering Science	=17	n=24
Geography	=2	n=4
Health Studies/Sciences	=9	n=11
Heritage Studies, Archival Studies	=1	n=1
Humanities	=18	n=50
Law	=10	n=18
Mathematics	=2	n=4
Medicine, Biomedical Engineering	=2	n=2
Media/Communications	=1	n=4
Natural Sciences	=20	n=29
Nursing, Midwifery	=7	n=9
Political Science	=3	n=6
Psychotherapy	=1	n=1
Social Sciences	=22	n=33

6.4.5.2 Likelihood of future engagement with web archives

Using a Likert scale for answer options, non-user participants (n=180) were asked about their likelihood of using the public NLI Web Archive in the future for their research, as well as some other online public web archives. Participants were provided with a list of six other online web archives (with a URL link to each resource). Table 6.12 provides a breakdown of responses.

Table 6.12: Representation of non-user responses (n=180) for the likelihood of future engagement with online public web archives

Participation per resource (n=180)	Definitely Likely	Fairly Likely	Unsure	Not Very Likely	Definitely Not Likely
NLI Web Archive	11.67% (=21)	30.00% (=54)	17.22% (=31)	28.33% (=51)	12.78% (=23)
Internet Archive, Wayback Machine	8.89% (=16)	26.67% (=48)	23.89% (=43)	23.89% (=43)	16.67% (=30)
PRONI Web Archive	3.89% (=7)	11.67% (=21)	20.56% (=37)	33.89% (=61)	30.00% (=54)
UK Web Archive	4.44% (=8)	24.44% (=44)	20.00% (=36)	27.22% (=49)	23.89% (=43)
UK Government Web Archive	2.78% (=5)	13.33% (=24)	20.00% (=36)	36.11% (=65)	27.78% (=50)
UK Parliament Web Archive	1.67% (=3)	11.11% (=20)	18.33% (=33)	33.89% (=61)	35.00% (=63)
US Library of Congress Web Archive	3.89% (=7)	16.11% (=29)	18.89% (=34)	32.22% (=58)	28.89% (=52)

From there, it is possible to calculate some measurements using filters, for a probability on whether awareness increases the likelihood of research engagement for each resource, by using the following formula.

Formula: number of participants who were unaware of a public web archive resource at the start of the survey, who also identified as a non-user*** and who specified as definitely likely* and fairly likely** to use the resource for future research (* + ** = [] ÷ *** x 100 = [] %).

Table 6.13 provides an overview of the application of the formula for each resource. It demonstrates a probability percentage, that awareness increases the likelihood for future research engagement with online public web archives for non-user respondents who were unaware of the existence of web archives prior to participation in the survey.

Table 6.13: Representation for the probability that awareness increases likelihood of engagement with online public web archives for non-users (n=180) who were unaware of the existence of online public web archives

Web archive resources	Unaware & non-user***	Definitely Likely*	Fairly Likely**	Increased Likelihood
NLI Web Archive	=162	=16	=46	38.27%
Internet Archive, Wayback Machine	=138	=7	=30	26.81%
PRONI Web Archive	=162	=5	=12	10.49%
UK Web Archive	=161	=6	=34	24.84%
UK Government Web Archive	=169	=4	=19	13.61%
UK Parliament Web Archive	=174	=3	=16	10.92%
US Library of Congress Web Archive	=163	=5	=18	14.11%

Participants were also asked about the likelihood that they would access or use a dark web archive in the future for their studies or research. They were informed that a dark web archive is only accessible onsite in a designated reading room or Library via an onsite portal. Figure 6.3 provides a breakdown of responses, calculated from participation in this section (n=180). As one can see, 5.00% (=9) of non-users responded with 'Definitively Likely', 12.22% (=22) with 'Fairly Likely', and 18.89% (=24) with 'Unsure'. While this seems like a low response towards the likelihood of using a dark web archive in the future, one needs to account that many of these respondents (=141 of 180) indicated their reasons for not using a web archive for research or study was due to a lack of awareness of their existence (see Table 6.10). Moreover, from the findings in an earlier question (section 4.4.3.1), of the total number of respondents in the survey, only 2% (4 of 239) indicated that they were aware of the existence of the NLI dark (.ie) web archive, meaning the concept of a dark (domain) web archive and its value as a research resource may not be clearly understood.

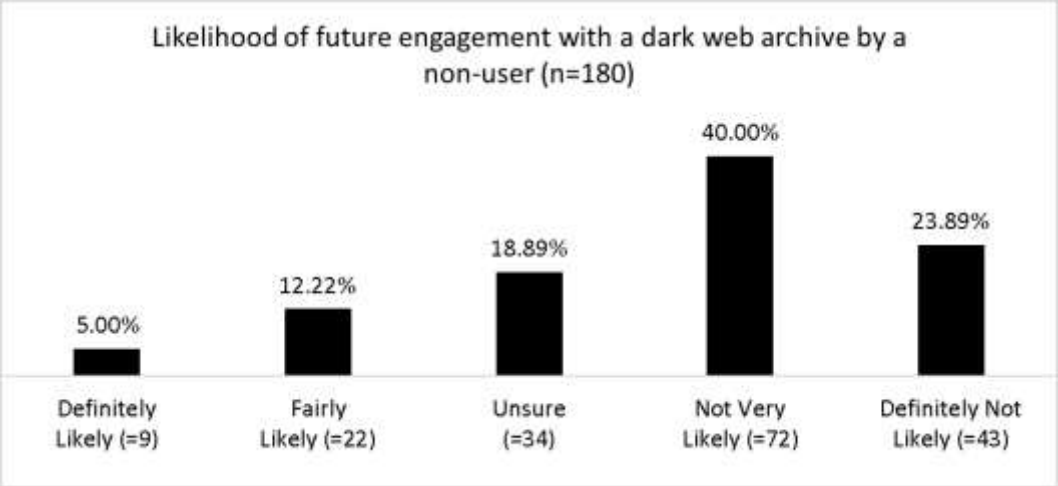

Figure 6.3: Representation for the likelihood of future engagement by a non-user (n=180) with a dark (domain) web archive

6.4.6 User Engagement with Web Archives

This section provides an overview of responses by user respondents (n=59) on their general reasons for using a web archive, their reasons for using a web archive for research, their use of online public web archives or dark web archives, and the likelihood they would use a dark web archive in the future.

6.4.6.1 Disciplines of user respondents

The discipline categories of user respondents (n=59) are outlined below in Table 6.14 and indicate that user respondents identify with a broad range of research fields.

6.4.6.2 General reasons for using a web archive

Respondents who identified as a user (n=59) were asked about their access or use of a web archive in general. Participants were provided with seven answer choices based on interests, along with an option of 'Other' to enter free text. They were asked to tick all that applied.

Figure 6.4 provides an overview of participant responses. It indicates that the vast majority (93.22%, =55) of user respondents utilise a web archive for research interests, followed by personal interests (72.88%, =43), historical interests (64.41%, =38), and cultural interests (42.37%, =25). One respondent chose the option for 'Other' and noted using a web archive to "recover old advertisements for teaching" (User, Business, Economics, Finance).

Table 6.14: Representation of discipline categories for user respondents (n=59)

Discipline Answer Choices		User (n=59)
1	Architecture	=1
2	Business/Economics/Finance	=3
3	Computer Science	=3
4	Dental Science	=1
5	Digital Arts/Humanities/Cultural Heritage	=1
6	Educational Science	=3
7	Engineering Science	=3
8	Geography (cartography, hydrology, meteorology, environment)	=2
9	Government/Public Administration	=1
10	Humanities (history, archaeology, languages, literature, philosophy, theology)	=21
11	Law (criminal, civil, common, statute)	=6
12	Mathematics	=1
13	Media/Communications	=3
14	Natural Sciences (biology, chemistry, physics, earth sciences, space sciences)	=2
15	Nursing/Midwifery	=1
16	Political Science	=2
17	Social Sciences (anthropology, human geography, linguistics, sociology, psychology)	=5

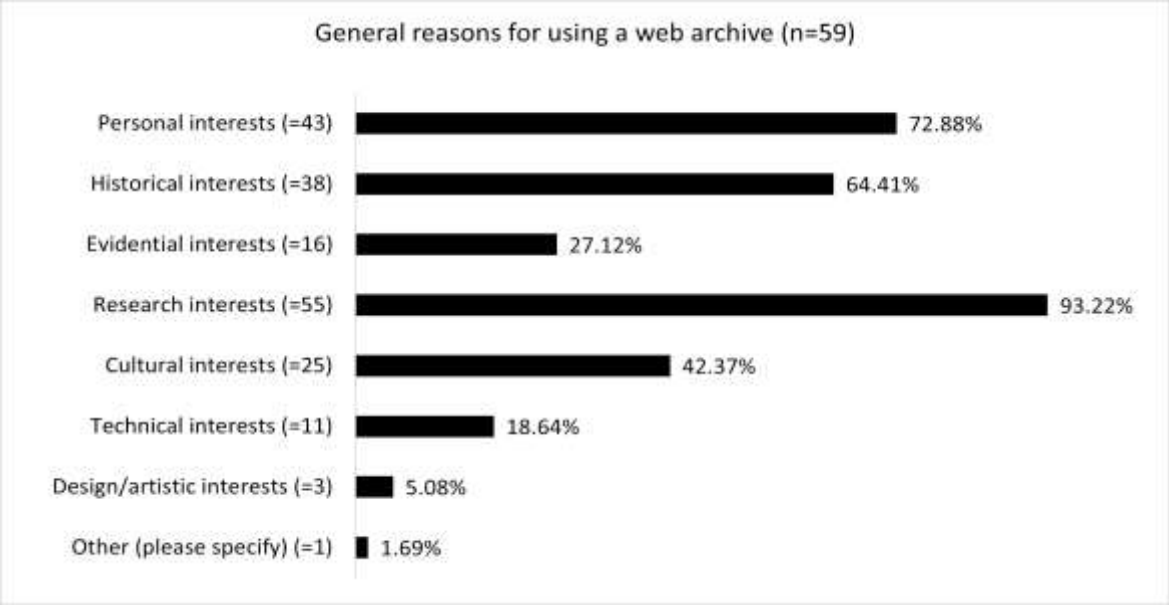

Figure 6.4: Representation of general reasons for using a web archive (n=59)

6.4.6.3 Reasons for using archived web content

User respondents (n=59) were asked about their use of archived web content (archived websites, blogs, web pages). Participants were provided with 6 answer choices, and an option for 'other reasons' to enter free text. They were asked to tick all that applied. Table 6.15 offers a breakdown of responses of participant reasons for using web archived content for their studies/research.

Further to this, 9 respondents entered free text which is summarised below.

- 3 respondents noted using a web archive for personal and historic interests .
- 1 respondent mentioned using archived web content as a secondary source for a thesis.
- 3 respondents referred to accessing content/websites no longer available on the web, with 1 respondent specifying "technical articles" (User, Engineering Science).
- 1 respondent mentioned access to "policy pages that were no longer publicly accessible, so as to compile evidence" (User, Humanities).
- 1 respondent used archived web content for a study of "longitudinal data concerning water parameters [as] sometimes websites only display current or recent results" (User, Dental Science) .

To break this down further, the reasons for using web archives for study or research by user respondents can be further organised by the following themes: for coursework purposes, for professional publication and historical research purposes; for teaching purposes; for

qualitative and quantitative research purpose; and for access to materials no longer available on the live web (Table 6.16).

Table 6.15: Representation of user participant reasons for using web archived content for their studies/research

Answer Choices	User Responses (n=59)
I have used archived web content as a primary source in an academic essay/assignment for my course	42.37% (=25)
I have used archived web content to document the history of an organisation in an academic essay/assignment for my course	20.34% (=12)
I have used archived web content as a primary source in a professional research report	16.95% (=10)
I have used archived web content as a primary source in a professional publication	27.12% (=16)
I have used archived web content to document the history of an organisation in a professional report/publication	11.86% (=7)
I have used archived web content as part of my teaching materials for undergraduate students	18.64% (=11)
I have used archived web content as part of my teaching materials for postgraduate students	22.03% (=13)
I have used large volumes of archived web content for content analysis / textual analysis / discourse analysis	8.47% (=5)
I have used large volumes of archived web content for data mining / topic modelling / data visualisation	6.78% (=4)
I have used large volumes of archived web content for network analysis / geo-spatial analysis	5.08% (=3)
I have used archived web content for other reasons not listed above - please specify	15.25% (=9)

Table 6.16: Representation of user respondent reasons for using web archives for study or research (n=59)

For coursework purposes (=38)	
● as a primary source in an academic essay/assignment for my course (=25) ● to document the history of an organisation in an academic essay/assignment for my course (=12) ● as a secondary source for a thesis (=1)	=38
For professional publication and historical research purposes (=33)	
● as a primary source in a professional research report (=10) ● as a primary source in a professional publication (=16) ● to document the history of an organisation in a professional report/publication (=7)	=33
For teaching purposes (=25)	
● as part of my teaching materials for undergraduate students (=11) ● as part of my teaching materials for postgraduate students (=13) ● to recover old advertisements for teaching (=1)	=25
For qualitative and quantitative research purposes (=12)	
● for content analysis / textual analysis / discourse analysis (=5) ● for data mining / topic modelling / data visualisation (=4) ● for network analysis / geo-spatial analysis (=3)	=12
For access to materials no longer available on the live web (=5)	
● for accessing content/websites no longer available on the web (=3) ● for access to "policy pages that were no longer publicly accessible, so as to compile evidence" (=1) ● for a study of "longitudinal data concerning water parameters" (=1)	=5

6.4.6.4 Use of online public web archives for studies/research

In terms of using online public web archive resources, user respondents (n=59) were asked if they ever accessed or used the online public NLI Web Archive and six other online public web archives for their studies or research. Table 6.17 provides a breakdown of responses by user respondents (n=59) for both questions and shows more than half of user respondents indicated being a user of the Internet Archive, Wayback Machine (=35). This was followed by the NLI Web Archive (=23), the UK Web Archive (=22), the UK Government Web Archive (=19) and the US Library of Congress Web Archive (=14). In examining the position and disciplines of respondents who use the public NLI Web Archive (=23 of 59), it reveals it is

used by educators (=9), students (=10) and researchers (=5) from nine different fields of research. A full breakdown of position representations, in line with discipline categories is available in Appendix G (Table G.1).

Table 6.17: Representation for the use of online public web archive by user respondents

Use of online public web archive resources	User Responses (n=59)
NLI Web Archive	=23
Internet Archive, Wayback Machine	=35
PRONI Web Archive	=5
UK Web Archive	=22
UK Government Web Archive	=19
UK Parliament Web Archive	=12
US Library of Congress Web Archive	=14

Participants were also asked if there were any other web archives that they accessed or used, and to name the resource as free text. 1 user respondent indicated 'Yes' and provided the following free text: "Unsure whether scientific journal and search engine archives count here -I use them" (User). One might consider that this respondent has a hazy understanding of the differences between a digital archive, a digital library, and a web archive, and this could indicate that the respondent may be confused about being a web archive user, and indeed, may not be a user at all. However, the respondent identified using the Wayback Machine "to access older versions of websites. I'm interested in longitudinal data concerning water parameters, but sometimes websites only display current or recent results" (User). Thus, it would be fair to suggest that while an individual might use a web archive for their research, they may still have a blurred understanding of the differences between a digital archive, a digital library, and a web archive.

6.4.6.5 Use of a dark web archive for studies/research

User respondents (n=59) were asked if they ever accessed or used a dark web archive for their studies/research. They were informed that a dark web archive is only accessible onsite in a designated reading room or Library via an onsite portal. Respondents were also asked to name the resources they used if they answered 'Yes'. 94.92% (=56) of user respondents

indicated 'No', and 3 respondents indicated 'Yes' but did not provide any free text to name the dark web archive they used.

6.4.6.6 Likelihood of future engagement by users with a dark web archive

Using a Likert scale, user respondents (n=59) were asked their opinions on the likelihood that they will access or use a dark web archive in the future for their studies/research. Figure 6.5 provides an overview of responses.

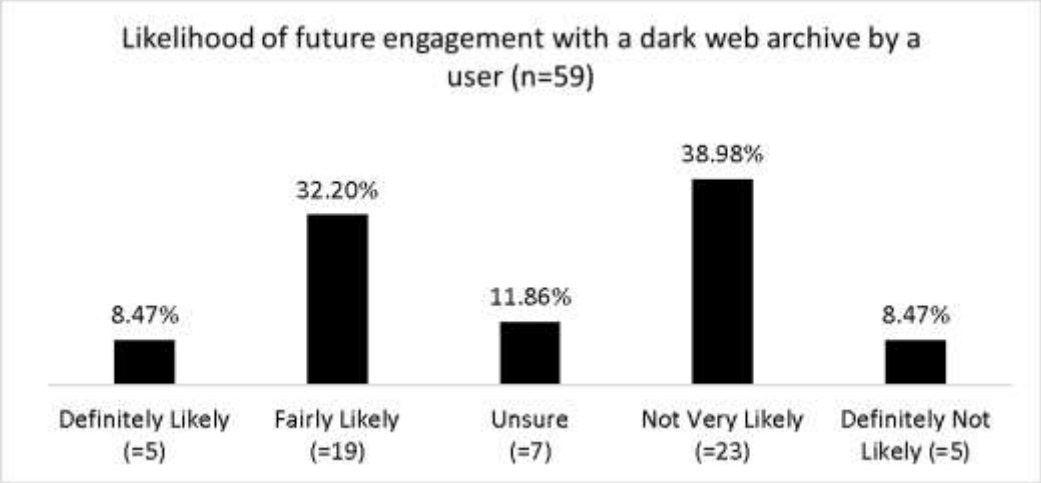

Figure 6.5: Representation for the likelihood of future engagement by users (n=59) with a dark web archive

As one can see, 8.47% (=5) of user respondents responded with 'Definitively Likely', 32.22% (=19) with 'Fairly Likely', and 11.86% (=7) with 'Unsure'. Compared with the non-user responses on the likelihood of using a dark web archive in the future (see Figure 6.3), user respondents offer a more positive outlook for the likelihood of using a dark web archive in the future. As already mentioned earlier, while the concept of a dark web archive, and its value as a research resource may not be clearly understood by many non-users, it also seems that some users may have a hazy comprehension of the concept of a dark web archive, its value as a research resource, and how it might be used. Earlier findings revealed that only 2% (4 of 239) of the total number of respondents in the survey were aware of the existence of the NLI dark (.ie) web archive (section 5.4.3). There is a need therefore, for a collaborative effort in raising awareness of its existence and to foster discussions regarding its future access and use for Irish based research.

6.4.7 Perceived value and importance of web archives

This section provides an overview of total participant responses (N=239) to questions on the perceived value of web archives, the importance of archiving websites and blogs based on topics, and whether web archives will become important as a resource for current, medium, or long-term future research in their field.

6.4.7.1 Perceived value of web archives

Using a Likert scale, and a list of six values, participants (N=239) were asked their opinion on the importance of archiving websites and blogs for current and future research, based on the values. Participants were offered a list of values with multiple choice options. Table 6.18 provides a full breakdown of responses. Over half of all respondents indicated 'Very Important' for historical value (63.60%, =152), followed by research value (55.65%, =133), and evidential value (53.97%, =129); and just under half of respondents indicated 'Very Important' for cultural value (48.95%, =117).

6.4.7.2 Perceived importance of archiving websites based on specific topics

Provided with a list of nine topics and a Likert scale, participants (N=239) were asked their opinion on the importance of archiving websites and blogs based on a topic area. Table 6.19 gives a breakdown of participant responses and shows that the most important topics to be archived are Direct Government websites, deemed as 'Very important' (64.46%, =166) and 'Fairly Important' (19.25%, =46); and Indirect Government websites, deemed as 'Very important' (58.16%, =139) and 'Fairly Important' (28.87%, =69). Moreover, the archiving of Science and Environment websites/blogs were rated as more important than the archiving of websites on Referendums, Politics, Elections and Events. This could be used as an indicator for future collection development policies, as the inclusion of such topics may appeal to a wider academic audience and thus, attract a broader range of engagement. It also emphasises the need for a more rigorous approach for the inclusion of direct and indirect governmental websites as part of national digital heritage.

Table 6.18: Representation of participant responses (N=239) for their perceived value of web archives

Participation per value (N=239)	Very important	Fairly important	Slightly Important	Not important	No opinion
Historical value	63.60% (=152)	22.59% (=54)	6.69% (=16)	1.67% (=4)	5.44% (=13)
Research value	55.65% (=133)	29.71% (=71)	5.86% (=14)	3.35% (=8)	5.44% (=13)
Evidential value	53.97% (=129)	25.94% (=62)	6.69% (=16)	3.77% (=9)	9.62% (=23)
Cultural value	48.95% (=117)	31.80% (=76)	10.88% (=26)	2.09% (=5)	6.28% (=15)
Technical value	25.52% (=61)	28.45% (=68)	25.10% (=60)	7.11% (=17)	13.81% (=33)
Design/artistic value	24.27% (=58)	24.69% (=59)	26.78% (=64)	11.30% (=27)	12.97% (=31)

Table 6.19: Representation of participant responses (N=239) on the importance of archiving websites/blogs based on topics

Participation per topic (N=239)	Very important	Fairly important	Slightly important	Not important	No Opinion
Direct Government (websites of official government departments or office holders, e.g., websites of the Department of Finance, the President, an Taoiseach)	69.46% (=166)	19.25% (=46)	4.18% (=10)	1.67% (=4)	5.44% (=13)
Indirect Government (websites of agencies deployed by the Irish Government to undertake a task, e.g., Irish Water, Nama)	58.16% (=139)	28.87% (=69)	5.86% (=14)	1.67% (=4)	5.44% (=13)
Politics (websites/blogs of political parties, political commentators)	50.63% (=121)	27.62% (=66)	10.88% (=26)	5.02% (=12)	5.86% (=14)
Community Groups/ Activists (websites/blogs of clubs, societies, advocacy groups, human rights groups)	42.26% (=101)	32.22% (=77)	13.39% (=32)	6.28% (=15)	5.86% (=14)

Events (websites/blogs for natural disasters, sporting events, commemoration events)	32.64% (=78)	38.49% (=92)	17.99% (=43)	4.60% (=11)	6.28% (=15)
Election Campaigns (websites/blogs of candidates, election judicators, commentators)	40.59% (=97)	30.54% (=73)	18.41% (=44)	5.02% (=12)	5.44% (=13)
Referendum Campaigns (websites/blogs of interest groups, referendum judicators, commentators)	46.86% (=112)	31.80% (=76)	12.55% (=30)	4.60% (=11)	4.18% (=10)
Environment (websites/blogs which report on climate change, pollution, conservation)	54.39% (=130)	31.38% (=75)	7.95% (=19)	2.93% (=7)	3.35% (=8)
Science (websites/blogs which report on advances in medicine, chemistry, physics)	57.32% (=137)	28.45% (=68)	6.28% (=15)	4.18% (=10)	3.77% (=9)

6.4.7.3 Importance of web archives as a resource for current, medium, or long-term future research

Using a Likert scale, participants (N=239) were asked their opinion on whether web archives would become important as a resource for current, medium, or long-term future research in their field. Table 6.20 offers a breakdown of participant responses.

Also, of interest are the discipline categories of respondents who indicated 'Yes', that web archives would become important.

- 98 respondents who indicated 'Yes' for current research identified with 20 discipline categories.
- 142 respondents who indicated 'Yes' for medium-term research identified with 20 discipline categories.
- 148 respondents who indicated 'Yes' for long-term research also identified with 20 discipline categories.

Table H.1 in Appendix H offers a full breakdown of discipline categories for the respondents mentioned above. Table 6.20 clearly indicates that many participants feel that web archives will become more important for research as time goes on, and so, there is a need to establish theoretical and methodological approaches to enable researchers and educators to work

with this type of data sooner rather than later. While Table H.1 demonstrates the need to consider potential research models and paradigms that are fit for purpose across a wide range of research fields.

Table 6.20: Representation of participant responses (N=239) on the importance of web archives for current, medium, or long-term future research

Participant responses on the importance of web archives for current, medium, or long-term future research	Yes	Maybe	No
Current research (next 5 years) (N=239)	41.00% (=98)	36.40% (=87)	22.59% (=54)
Medium-term research (5-15 years) (N=239)	59.41% (=142)	29.71% (=71)	10.88% (=26)
Long-term research (15+ years) (N=239)	61.92% (=148)	27.62% (=66)	10.46% (=25)

6.4.8 Perceived challenges for the use of archived web content for studies/research in the future

Finally, at the end of the survey, participants were provided with an optional open-ended question, and asked their opinion on their perceived challenges for the future use of archived web content in their field of research. 49 respondents entered free text of which 14 identified as a user, and 35 as a non-user. The free text was coded through the number of times a particular challenge was mentioned in participants' answers. For example, one participant may mention multiple challenges in one response, and thus, each individual challenge mentioned is included as a representation (R/r=).

Table 6.21 offers a breakdown of the thematic representations of responses by participants (n=50) on their perceived challenges for the future use of archived web content in their field of research and is organised into four main sub-themes as follows:

- Using web archives and archived web content (r=40)
- Awareness of the existence, content and value of web archives (r=16)
- Data management and preservation (r=11)
- Not relevant for research topic (r=4)

These sub-themes are discussed in more detail in the next section.

Table 6.21: Thematic representation of participant responses on their perceived challenges for the future use of archived web content in their field of research (n=50)

Theme representation of responses for perceived challenges for the future use of archived web content (n=50)	No. of coded representations (R=60)
|> Using web archives and archived web content ● Search and navigation (r=10) ● Volume of data (r=10) ● Access and discovery (r=9) ● Representativeness and completeness of the data (r=5) ● Non-established source/source credibility (r=4) ● Citation (r=2)	r=40
|> Awareness ● Awareness of the existence, content and value of web archives (r=16)	r=16
|> Data management and preservation ● Data management and data reliability (r=5) ● Storage (r=3) ● Technical challenges (r=3)	r=11
|> Not relevant for research topic	r=4

6.4.8.1 Using web archives and archived web content

The responses presented several representations on the challenges for using web archives and archived web content in different contexts (r=40). These representations are further broken down into the sub-themes below.

- Search and navigation (r=10)
- Volume of data (r=10)
- Access and discovery (r=9)
- Representativeness and completeness of the data (r=5)
- Non-established source/source credibility (r=4)
- Citation (r=2)

Search and navigation

10 representations refer to challenges on how to use a web archive in terms of search and navigation. Some examples are outlined below.

- "Searching through the mass of material without a traditional kind of curated catalogue" (User, Social Sciences)
- "Individual researcher knowledge about how to use search functions on web archives" (Non-user, Natural Sciences)
- "searchability of archived data" (Non-user, Engineering Science)
- "Techniques for searching" (Non-user, Social Sciences)
- "navigation hopefully by content indexing" (User, Computer Science)
- "how to [...] search through it" (User, Social Sciences)
- "how to navigate them" (Non-user, Humanities)

Volume of data

10 representations mention challenges in dealing with large volumes of data. Out of this, 3 representations refer to big data analytics, but from different perspectives. 1 representation notes a lack of training for humanities researchers using large scale data, while 2 representations show some concern about the use of big data analytics. Some examples of these representations are provided below.

- "Weeding the wheat from the chaff due to the sheer volume of information" (Non-user, Social Sciences)
- "The very large quantity of material" (Non-user, Humanities)
- "The volume of material available for a single organisation or an event (e.g. an election) may exceed the volume available for similar events or organisations in the past. Researchers will therefore have to deal with a far greater level of data than their predecessors" (User, Humanities)
- "The scale of the data available and the current deficit of training in Ireland in tools for large scale data analysis for humanities researchers." (User, Humanities)
- "I would be concerned that the increasing use of data mining and other techniques to analyse large volumes of such content will result in only a partial analysis/understanding of any topic as the absence/balance of material may be misunderstood/misinterpreted." (User, Dental Science)
- "The quantity of information available and the difficulty of understanding reception and audience. In some senses these are the same problems encountered by

historians of the 20th C already in terms of print but they are on a very large scale. There may be a temptation to move towards increased emphasis on quantification thus losing some of the nuance and interest of qualitative approaches." (Non-user, Humanities)

Access and discovery

9 representations refer to accessibility, access, or discovery with 1 user respondent specifically mentioning access to the NLI domain web archive. Another representation suggests that web archives will be underutilised if they are not findable through search engines, especially for those who are unaware of their existence. Some examples of these representations are outlined below.

- "Ensuring open access regardless of location e.g. if the National Library makes its [domain] web archive access [...] in a single location this would probably be in Dublin and non-Dublin based researchers would then have very limited access" (User, Engineering Science)
- "how to access them" (Non-user, Educational Science)
- "The accessibility [...] of archived data" (Non-user, Engineering Science)
- "Access and technological limitations" (Non-user, Computer Science)
- "Are these archives harvested by engines like Google? If not, they may be under-utilised as people who are not aware of them may not go directly to the archive to search for relevant material" (User, Humanities)
- "accessibility" (User, Humanities)
- "how to access them (Non-user, Educational Science)

Representativeness and completeness of the data:

There are 4 representations which mention the representativeness of the data in web archives, in terms of what is presented (or not presented) on the web and what ends up in a web archive. Examples are provided below.

- "My specific topic of research is controversial and scientific/ professional/commercial organisations tend to keep a minimum of information about it online. On the other hand, political/community/environmental groups with opposing views are vocal and prolific online. In my experience, research using archived web content will be constrained by what does and does not become web content, and these decisions and the reasons for them may not be evident from a simple examination of the content that is available. In other words, the context in

which the content is created could be lost over time, even if the content itself is preserved." (User, Dental Science)

- "The representativeness of the archived web content" (Non-user, Health Studies/Sciences)
- "Determining how representative the [...] content is for the general public or particular groups and their actual opinions, behaviors, and/or thoughts" (User, Business, Economics, Finance)

1 representation notes the completeness of the data due to chronological gaps in capture dates

- "In the case of the Wayback Machine, the unpredictable (and occasionally somewhat erratic) frequency of archiving can be problematic - you might have several snapshots in a month, and then miss all of the following year." (User, Media/Communications)

Non-established source and source credibility

4 representations mention, in some way, that the use of web archives and archived web content is not an established or credible source for research. Some examples are provided below.

- "most students and researchers as far as I know tend to use more traditional sources (books, journals, patents) and since this is an established method of research it will be hard to break." (Non-user, Engineering Science)
- "It is a largely unknown entity. It might be difficult to convince my supervisors that it is a valid source of credible information." (Non-user, Social Sciences)
- "The credibility of such sources may be raised... Questions around why they are no longer active etc which likely have perfectly reasonable reasons but could cause doubt" (Non-user, Heritage Studies, Archival Studies)
- "The content of a web archive would not be a 'credible' citation in the legal field" (Non-user, Law)

Citation

2 representations note citation as a challenge for the use of archived web content.

- "researcher knowledge of how to cite archives and versions of historical web pages viewed on archives so that research is reproducible." (Non-user, Natural Sciences)
- "Citation systems" (Non-user, Humanities)

6.4.8.2 Awareness of the existence, content, and value of web archives

In terms of challenges for future engagement, there are 16 representations which refer to awareness in the context of awareness of the existence, the content, and the value of web archives for research. Several examples are outlined below.

- "How to increase awareness of their existence" (Non-user, Lecturer, Humanities)
- "public awareness of the archives" (User, Humanities)
- "Making people aware of such archives both within university/college information systems and in public libraries/information kiosks/public service web portals" (Non-user, Computer Science)
- "It is a largely unknown entity" (Non-user, Social Sciences)
- "How to increase awareness of their existence" (Non-user, Humanities)
- "Aware of the content" (Non-User, Educational Science)
- "Knowing that they exist" (Non-user, Architecture)
- "knowing where and what is there" (Non-user, Medicine, Biomedical Engineering)
- "Making researchers aware of relevant and informed content" (Non-user, Engineering Science)
- "Demonstrating value to future users as there is a habit now of searching the web for everything [...] You need to show people papers, projects where the value of an archive is demonstrable or at some type of culture night thing or something so people learn from such archives" (Non-User, Computer Science)

6.4.8.3 Data management and preservation

5 representations mention data management, reliability, storage, or preservation in some context. Some examples are provided below.

- "proper and appropriate maintenance of the content" (Non-user, Professor or Associate Professor, Humanities)
- "Data management of these research resources. Loss of important documents like email communications will be a factor too re non-archiving of these tech files " (Non-user, Researcher, Humanities)
- "Ensuring that the content has not been tampered with" (User, Humanities)
- "cost of storage" (Non-user, Built Environment)
- "Storage capacity" (User, Computer Science)

3 representations note technical challenges in terms of the capture and preservation of web content, and the provision of access to archived content in the context of the changing nature of browsers, platforms, and software. Examples are provided below.

- “I study and research social (digital) media and new apps and platforms are constantly being developed, while others close or become defunct. Archiving content on closed or defunct sites will be difficult if/when the technology advances to the point where these are no longer compatible with the latest operating systems, browsers, or mobile devices.” (User, Media/Communications)
- “technical access (risks associated with obsolescence of software or devices)” (Non-user, Humanities)
- “Access and technological limitations (e.g. browser support for older technology, etc)” (Non-user, Computer Science)

6.4.8.4 Not relevant for research topic

4 representations infer that the lack of use of archived web content in their field of study was due to the non-relevance of a web archive for their research. Some examples are provided below.

- “My field is not directly served by the above categories (ancient history).” (Non-user, Humanities)
- “Important data will remain online. There will be no need for ‘wasting’ time digging in old files - I am talking about science. For humanities, it might differ.” (Non-user, Natural Sciences)
- “I just don't think they're relevant to science. We access peer-reviewed scientific literature. Web archives have a place, but not in my research.” (Non-user, Natural Sciences)

6.5 Discussion

The survey results and analysis are based on a final number of 239 respondents, of which 59 respondents identified as a user (24.69%), and 180 respondents identified as a non-user (75.31%). The terms user and non-user relate to whether a respondent specified that they have used or not used a web archive for their research or studies. The survey collected enough quantitative data combined with an element of qualitative data to provide useful insights for a discussion to address some of the research questions proposed for this chapter. First, the findings show that a large majority of all respondents (N=239) use the web for research, of which 88.28% of respondents indicated ‘Always’, and 11.30% indicated

'Sometimes', demonstrating that the web is a major research resource in Irish third-level academic institutions (see Table 6.7).

6.5.1 Current level of awareness for the existence of web archives

Respondent (N=239) awareness of online public web archives differed, of which the Internet Archive's Wayback Machine was the most widely known resource (31.38%), followed by the US Library of Congress Web Archive (23.85%), and the UK Web Archive (21.76%). In addition, 13.81 % (n=33) of respondents indicated awareness of the PRONI web archive, and 18.41% (n=44) of respondents indicated awareness of the public NLI Web Archive, and identified as educators, students, and researchers, from 11 different discipline categories.[55] Awareness of the online public NLI Web Archive vis-à-vis the rate of respondents per discipline category was highest for the Humanities; however, it was quite low for other discipline categories such as the Social Sciences, Engineering Science, and the Natural Sciences. While there is a large gap between the awareness of the Wayback Machine (31%), and both the NLI Web Archive (18%) and PRONI web archive (13.8%), Riley and Crookston (2015) also found a gap in their New Zealand study in so far as much more respondents were aware of the Wayback Machine than were aware of the New Zealand Web Archive (p. 12). Riley and Crookston (2015) submit that this may be due to the high profile of the Wayback Machine, poor efforts to promote the New Zealand Web Archive, and because their web archive collections are not in a standalone resource, rather they are integrated in a common interface which includes other library collections (p. 12).

In the case of the online NLI Web Archive and PRONI Web Archive, the gap may similarly be due to a lack of promotion but could also be due to a lack of use cases in Irish third-level education and research, which might showcase the use of the resources and thus encourage more use. Moreover, the NLI only began a web archiving initiative in 2011, and PRONI in 2010, and as such they are relatively young archives in comparison to some others. Therefore, it is encouraging to see some levels of awareness for these archives, which can be built upon for further promotion, outreach, and collaboration.

On the other hand, awareness of the NLI domain web archive is quite poor at 2%, and thus will warrant a strategy for promotion as a research resource, when it eventually becomes accessible. In this regard it will be essential for the NLI to be afforded the capacity to collaborate with users and promote the resource to potential users; and to build solid

[55] *Business, Economics, Finance; *Computer Science; *Digital Arts/Humanities/Heritage; *Educational Science; *Engineering Science; *Geography; *Humanities; *Law; *Media/Communications; *Natural Sciences; *Social Sciences

research infrastructures between the NLI web archive, and the research teams seeking to use the data (Brügger, 2021c). This will require funding, and a cultural shift placing the creator and user as partners in the full web archiving lifecycle.

In addition, as noted by one respondent, access to a domain archive onsite in the NLI reading room only will present a geographical barrier for some researchers. Maurer (2022) also discusses how the provision of onsite 'only' access to web archive collections makes them geographically inaccessible for many researchers. Chapter 4.0 also showed how onsite access may present barriers for engagement due to socio-economic reasons (section 3.5.7). Therefore, in terms of the establishment of an Irish domain web archive, the obvious solution to the access problem would be to make it open access using an 'Opt-Out' strategy. However, this is probably unlikely for all types of web content. Therefore, for content that requires restrictions, such as content behind paywalls, there will be a need to consider how access can be provided in more than one geographic location, perhaps in conjunction with other legal deposit libraries across Ireland. Moreover, access provisions should be made for researchers and users who are not affiliated to an academic institution. In the long-term, access should be provided in public libraries across Ireland, and this would ensure that users are not disadvantaged based on geographic location or socio-economic circumstances.

It must also be emphasised that certain categories of websites should be open access by default, including:

i. websites belonging to the Irish government, its departments, and its subsidiary agencies, as well as local government and councils,
ii. websites belonging to public bodies, quangos, civic agencies, and political parties who receive government funding in any form,
iii. websites belonging to owners or organisations who have received funding from the Irish government or any of its subsidiary agencies, and this should be stipulated as part of any funding agreement, and
iv. websites which have a variety of Creative Commons licences could also be considered for inclusion for open access.

6.5.2 Terminology

As mentioned in section 6.3.4, the data from 46 survey participants was not included in the final analysis of this survey study; nonetheless, it still warrants inclusion in the discussion regarding awareness of web archives as resources for research. There is ample evidence from the 46 submissions to suggest that respondents were confused as to what a web archive is and, for the most part, respondents correlated the meaning of a web archive to

that of a digital library, digital archive, or digital data repository. There was a similar occurrence of this in the WARST survey (section 4.3.4), corresponding with the observations of Costea (2018) 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" (p. 11). Brügger (2018) also highlights the difficulty with the term. He discusses whether a web archive best fits to the family of an archive or library, but notes that while they may be misleading, the terms web archive and web archiving were coined decades ago and are part of the vernacular for this type of resource (pp. 77–78).

6.5.3 Reasons for a lack of engagement with web archives for research

Of total respondents (N=239), 75.31% (n=180) acknowledged that they did not engage with web archives for their research or studies. The reasons for this are varied; however, a large majority of non-user respondents (141 of 180) indicated that non-engagement was due to a lack of awareness of the availability of web archives as resources for research. Furthermore, the findings suggest that the value of web archives as a research resource is not clearly understood by an unfamiliar audience. For some respondents, there is a need for more efforts to demonstrate the importance of archiving the web and to promote the value of web archives for research. Therefore, it could be surmised that there is a need for the dissemination of use cases in Irish based research that will demonstrate the use of web archives as a research resource. These findings compare well with the findings from other user studies, whereby Jatowt et al (2008), Riley and Crookston (2015), and Costea (2018) also found a large lack of awareness of the existence of web archives by the participants in their studies. Winters (2017) also points to a lack of awareness as being one of the major reasons as to why web archives are not being more utilised (p. 174).

Challenges in using a web archive, and web archive content, for research were also mentioned by non-user respondents whereby 81 participants did not know how to find archived websites relevant to their studies/research in a web archive, and 75 participants did not know how to use a web archive for their studies/research. While Costea's (2018) study concludes that a lack of scholarly use of web archives is related to a lack of awareness of their existence, Costea also notes that many researchers were unaware of the content of a web archive, and how a web archive can be used as a resource for research (p. 25). Thus, the findings of this chapter also correlate to the findings by Costea (2018).

In terms of other reasons for a lack of engagement, 47 participants were unsure of the credibility or authority of using archived websites, which demonstrates pedagogical issues. In addition, 32 participants did not know how to cite/reference an archived website which

correlates well with the findings from section 3.5.5, whereby individuals from both the web archiving community and the scholarly user community experienced challenges for the citation of archived websites or derived datasets of archived web content. 24 participants indicated that they were unsure about copyright implications for using archived web content. Interestingly only 14 respondents felt that they did not have the technical skills to use a web archive for their studies/research, which seems quite low. However, one should consider here that the respondents are non-users and therefore, have yet to discover the various types of technical skills and tools which are required for participation in web archive research as a user and were discussed in chapter 3.0 (section 3.5.3 & section 3.5.6).

6.5.4 Likelihood of a non-user using a web archive for research, after becoming aware of its existence

There is good reason to believe that creating awareness of the availability of online public web archives increases the likelihood of researcher engagement. With the use of filters, this chapter demonstrates that awareness increases the probable likelihood of research engagement by non-users (n=180), with percentages of 44.40% for the NLI Web Archive, 26.81% for the Wayback Machine, and 24.84% for the UK Web Archive. However, while the findings suggest that awareness increases the likelihood for an increase in researcher engagement, there is an indication that the promotion of the existence of web archives by itself may not be enough. For example, for some respondents, more efforts are needed to demonstrate the research value of web archives. Again, this highlights a need for more use cases in Irish based research to demonstrate approaches for using web archives as a resource, and the need for research infrastructures between web archive creators and web archive users to assist in promoting the value and use of these resources. In terms of the likelihood of a non-user respondent using a dark (domain) web archive for their research in the future, the response rate is low. However, it is also suggested that one needs to account for the fact that many of these respondents indicated that they did not use web archives for their research due to a lack of awareness of their existence (see Table 6.10). Thus, it could simply be a case that the concept of a domain web archive, and its value as a research resource, is not clearly understood.

6.5.5 Challenges perceived by scholars for the future use of archived web content

Of all participants (N=239), 50 respondents provided text responses for the challenges they perceived for future engagement with archived web content in their research fields. The text was analysed and broken down into four main themes, and sub-themes as outlined below.

- Using web archives and archived web content
    - Search and navigation
    - Volume of data
    - Access and discovery
    - Representativeness and completeness of the data
    - Non-established source/source credibility
    - Citation
- Awareness of the existence, content, and value of web archives
- Data management and preservation
    - Data management and data reliability
    - Storage
    - Technical challenges
- Not relevant for research topic

The perceived challenges presented by respondents are certainly useful in understanding how we might proceed to incorporate the use of web archives for teaching and for conducting Irish based research, alongside more traditional sources and methods. Of interest are the different outlooks on the use of large-scale analysis. 1 respondent notes the need for training in big data analysis for Humanities, while 2 respondents are concerned that big data analysis does not account for a full understanding of the context of the data. Rather, this might be better achieved with a qualitative approach. This implies that there is a need to consider research models that consider both qualitative and quantitative methods as standalone practices, or a mixture of both as a combined approach to include web archives as a resource for research in Ireland.

6.5.6 Users of web archives in Irish academic institutions

The survey results did not present any significant patterns to suggest that nationality, age, or gender have any influence on engagement with web archives. Of respondents who identified as a user (n=59), 30 respondents identified as educators, 19 as students and 10 as researchers. The data also shows that user respondents identified with 17 discipline categories.[56] As there has been a recent growth in the literature which promotes the use of web archives as resources for research in the humanities and social sciences (Gomes et al.,

[56] Architecture; Business/Economics/Finance; Computer Science; Dental Science; Digital Arts/Humanities/Cultural Heritage; Educational Science; Engineering Science; Geography; Government/Public Administration; Humanities; Law; Mathematics; Media/Communications; Natural Sciences; Nursing/Midwifery; Political Science; Social Sciences

2021b; Brügger & Laursen, 2019; Brügger & Milligan, 2019; Brügger, 2018; Milligan, 2019; Brügger & Schroeder, 2017; Ogden, 2022; Gorsky, 2015), it is perhaps no surprise to see a strong number of users from the 'Humanities' in this study. On the other hand, there was a low-level of users from the 'Social Sciences'. The findings show that this is most likely due to a lack of awareness of the existence of web archives.

In terms of using a web archive in general, a large majority of user respondents (n=59) indicated that they use web archives for research interests (93.22%, =55). Respondents also indicated the use of a web archive for personal interests (72.88%, =43), historical interests (64.41%, =38) and cultural interests (42.37%, =25).

User respondent reasons for using archived web content for their studies or research are further outlined in Table 6.22. It indicates that user respondents utilise web archives and archived web content for coursework purposes, for professional publication and historical research purposes, for teaching purposes, for qualitative and quantitative research purposes and for access to materials no longer available on the live web. What is also surprising is that users come from a diverse range of research fields, which reflects that both multidisciplinary and interdisciplinary deliberation are required to consider the challenges, and potential solutions, for developing research models and paradigms for the use of web archives for Irish based research that are fit for purpose in a broad spectrum of research fields. Stember's (1991) description of the terms multidisciplinary and interdisciplinary is a useful guide here. For Stember, (1991) multidisciplinary entails a collaboration between individuals from different disciplines "who each provide a different perspective on a problem or issue", and interdisciplinary is a step up from that to entail a collaboration between individuals from different disciplines to integrate methods and knowledge "into harmonious relationships" through a synthesis of strategies and approaches (p. 4).

Table 6.22: Combined data from Section 3.6 for user participant (n=59) reasons for using archived web content for their studies or research

For coursework purposes
● as a primary source in an academic essay/assignment for my course ● to document the history of an organisation in an academic essay/assignment for my course ● as a secondary source for a thesis
For professional publication and historical research purposes
● as a primary source in a professional research report ● as a primary source in a professional publication ● to document the history of an organisation in a professional report/publication
For teaching purposes
● as part of teaching materials for undergraduate students ● as part of teaching materials for postgraduate students ● to recover old advertisements for teaching
For qualitative and quantitative research purposes
● for content analysis / textual analysis / discourse analysis ● for data mining / topic modelling / data visualisation ● for data mining / topic modelling / data visualisation ● for network analysis / geo-spatial analysis
For access to materials no longer available on the live web
● for accessing content / websites no longer available on the web ● for access to "policy pages that were no longer publicly accessible, so as to compile evidence" ● for a study of "longitudinal data concerning water parameters"

Certainly, the user responses in this study offer some valuable insights on the opportunities for the use of web archives for Irish based research, and there is reason to believe that this community will grow over the next few years, as more academics become aware of web archives as resources for research. However, increases in web archive engagement will also depend on the promotion of awareness of the value of web archives, and demonstrations of use cases in academia as well as the public sphere. Formulating an Irish based multidisciplinary/interdisciplinary research network for current scholarly users and potential users, as well as web archivists, information professionals and technicians and web design professionals, would be of great benefit here. It would assist in addressing potential

solutions for developing research models and paradigms for the use of web archives for Irish based research that are fit for purpose in a broad spectrum of research fields. It would further enable discussions to develop frameworks to provide course modules for students in the use of web archives for research, and training courses for educators on how to incorporate web archived content as part of their teaching materials and methods.

6.5.7 Perceived importance of archiving websites based on specific topics

Regarding the perceived importance of archiving websites based on a topic area, participants (N=239) considered Direct Government websites and Indirect Government to be of highest importance. Of further interest, respondents rated the archiving of Science websites and Environment websites as more important than websites on Politics, Referendums, Elections and Events. This could be used as an indicator for future collection development policies, as the inclusion of such topics may appeal to a wider academic audience and thus attract a broader range of engagement. Moreover, it emphasises the need for a more rigorous approach for the inclusion of direct and indirect governmental websites as part of national digital heritage.

6.5.8 Perceived value of web archives

Of total participant responses (N=239), 63.60% of respondents perceived the historical value of a web archive to be 'Very Important', followed by research value (55.65%), evidential value (53.97%) and cultural value (48.95%). Regarding whether web archives would become important as a resource for current, medium, or long-term future research in their field, many participants indicated that web archives will become more important for research as time goes on. So, there is a need to establish theoretical and methodological approaches, to enable researchers and educators to work with this type of data sooner rather than later. Furthermore, these participants identified with 20 disciplines, again demonstrating the need to consider potential research models and paradigms that are fit for purpose across a wide range of research fields.

6.6 Summary

This chapter set out to provide some insight into the awareness of, and engagement with, web archives in Irish third-level academic institutions to gain a better understanding of how and why archived web content is used or not used for research in Ireland. The chapter was also exploratory in terms of assessing some of the opportunities and challenges for using web archives, and considerations for how to best facilitate their use, going forward. The

chapter engaged with desk research, a review of related literature, and an online survey of lecturers, students and researchers based in Irish academic institutions to examine levels of awareness of, and engagement (or non-engagement) with web archives in Irish academic institutions (RQ4). It further examined some of the perceived challenges by Irish based researchers for the future use of web archives for their research or study (RQ2) and offered some perspectives on approaches for improving the conditions for conducting web archive research (RQ5).

The chapter discussed the current level of awareness for the existence of web archives, the challenges with terminology, the reasons for a lack of engagement with web archives for research, and the likelihood of a non-user using a web archive for research after becoming aware of its existence. It also examined the users of web archives, the use of web archives and archived web content for research, and the challenges perceived by Irish based researchers and students for the future use of archived web content. Participants also gave their opinions on the perceived importance of archiving websites based on specific topics, and the perceived value of web archives.

From the foregoing, the findings demonstrate a limited awareness of the existence of web archives in Irish academic institutions, and, for an unfamiliar audience, more effort is needed to demonstrate the importance of archiving the web and to promote the value of web archives as resources for research. On the other hand, the findings revealed that there is a small community of web archive users in Irish academic institutions, at different levels of education and academia, aged from 18 to 65 years, and from a broad range of research fields. So, there is already a starting base of scholarly users and potential users which could be built upon to promulgate discourse for developing multidisciplinary and interdisciplinary research networks with web archivists, information professionals and technicians, as well as web design professionals, to address potential solutions for developing research models and paradigms for the use of web archives for Irish based research that are fit for purpose in a broad spectrum of research fields. It would also enable discussions to develop frameworks to provide course modules for students in the use of web archives for research, and training courses for educators on how to incorporate web archived content as part of their teaching materials and methods.

7.0 THE FUTURE(S) OF WEB ARCHIVE RESEARCH

This thesis aimed to investigate web archive research in Ireland in line with international developments, through the integration of desk research, survey studies, and case studies, using a combination of research methods, qualitative and quantitative, drawn from multiple discipline areas. In a broad sense, the thesis may be positioned at the intersections of the humanities and information science, and engages with scholarship and perspectives from archival science, library and information science, heritage studies, computer science, social sciences, media studies, cultural studies, humanities studies, and the evolving field of web archive research. From the outset, the thesis positioned web archive research to be inclusive of the processes and activities described in the Archive-It web archiving lifecycle model which includes appraisal, selection, capture, storage, quality assurance, preservation and maintenance, replay/playback, access, use and reuse (Bragg & Hanna, 2013).

First, the thesis positioned heritage within the broader framework of societies, communications, culture, and argued that it is within the intersections of these concepts that heritage is produced. In doing so, it provided an understanding of a society as a large social group, made up of individuals who interact and communicate, who have multiple things in common (e.g., territory, language, traditions, culture, political institutions), and have some levels of consciousness that they differ from other societies. It further offered some insights on what constitutes the heritage of a society, and how national heritage should be inclusive of a society's sub-groups, ethnic groups, and communities. The thesis demonstrated how legal nationality as citizenship does not ensure affiliation to the nation of the state. Moreover, one should also consider that since the conception of the Irish Free State up until the Millennium, the social, cultural, and political institutions of the Republic of Ireland have been accustomed to catering for a nation of settled white Irish Catholics (Howard, 2016). Thus, it could be argued that this will have an influence on the production and preservation of Irish national heritage.

Therefore, the thesis suggests that there will always be a need to consider how collection development policies for national heritage collections may revolve around a dominant hegemonic social group, at the cost of excluding representations from ethnic minorities, societal sub-groups, or alternative communities. Moreover, when it comes to the preservation of national digital heritage on the web and social media, there needs to be ongoing discussion on what gets captured and why, and how it reflects the ongoing transformations in societies, communications, and culture. The thesis further demonstrated

how there is a need for ongoing discussion on what constitutes national heritage in line with the politics of representation, preferred meanings, and alternative discourses, how these concepts influence what is included or excluded in the preservation of national heritage, and/or how this translates into the gaps and silences which are inherent in national archives and libraries.

Thereafter, the thesis explored some of the underlying reasons for web archiving, and how these stemmed from wider concerns on the loss of digital heritage in general, with the web just being another media carrier to worry about. The thesis examined the causes for the loss of digital heritage, inclusive of Irish digital heritage, and explored the challenges for participation in web archive research, and how this relates to Ireland. Through a survey study the thesis examined the challenges for the creation and use of web archives, and the overlaps, and intersections of such challenges across communities of practice within web archive research. The thesis assessed the usefulness of web archives based in Ireland, and their availability, and accessibility as resources for conducting research on Irish based topics. And, through another survey, the thesis investigated the awareness of, and engagement (and non-engagement) with, web archives as resources for research in Irish academic institutions.

For clarity, the thesis referred to Irish digital heritage in the context of the digital heritage of the island of Ireland, and when required, it referred to the digital heritage of Northern Ireland (NI) or the Republic of Ireland (ROI) to distinguish between the two jurisdictions.

This chapter acts as a final summation of the thesis, and will revisit the research problem, and the research questions and answers through a synthesis of the findings and discussion from the various chapters. The chapter concludes with some final thoughts from the researcher, and some suggestions for future work.

7.1 Revisiting the Research Problem

Since its invention in the early 1990s, the web has become a major resource for researchers. Yet it is a transient medium: information is in constant flux with content removal and updates, and the omnipresent '404 Not Found' error. As the early web materialised, concerns about the ephemeral nature of the web also emerged. National libraries and cultural heritage organisations soon realised the need to preserve informational content on the web and the development of web crawlers gave rise to the technology for archiving the web. It is widely agreed that web archiving involves the selection and collection of web content, preserving it for the future and making the collected web content available for

access and use. However, concerns about the ephemeral nature of the web also stemmed from already existing apprehensions regarding the storage and preservation of computational records, electronic information, multimedia and born digital materials in general (Fishbein, 1972; Dollar, 1978; Committee on the Records of Government, 1985; Graham, 1994; Waters & Garrett, 1996; Gardner, 1997; Kuny, 1997). From the 1990s, many western societies undertook a review of their copyright, heritage, and legal deposit laws, due to the developments in electronic/online publishing; and many have implemented reforms to account for the legal deposit of non-print materials, including the development of national web archiving programmes.

As more of the cultural, historical, legal, evidential, informational, and social record happens on the web, heritage institutions are tasked with keeping up with ongoing technological changes to capture and preserve this transient medium. Because of the enormity of the task, it is at least unreasonable, and probably impossible, to expect any one institution to assume full responsibility for archiving everything on the web. Therefore, a multi-agency worldwide approach (mostly in developed countries) has materialised whereby different institutions in different countries endeavour to preserve what they can, and what they deem as relevant for their mandate and stakeholders. In the interests of national heritage, the onus is often on national libraries to save what they can of the national web space inclusive of a selective permissions-based approach, and the routine archiving of the national web domain.

While it may seem inevitable that researchers in the humanities, media studies, and social sciences will integrate archived web content with more traditional formats for research topics from the mid-1990s, scholars have been slow to engage with web archives as resources for research (Webster, 2020; Rogers, 2019; Leetaru, 2019; Meyer et al., 2017; Webster, 2017b; Winters, 2017; Leetaru, 2017; Brügger, 2016; Meyer et al., 2011; Dougherty et al., 2010). In Ireland, the publication of Irish based research integrating the use of archived web content is difficult to find with a few exceptions being Malone (n.d.), Harjani (2018), Byrne (2019), Webster (2019), Greene & Ryan (2019), and Greene (2020). Prior to conducting the research for this thesis, very little was known about the scholarly awareness of web archives, and the reasons for engagement or non-engagement with web archives in Irish academia.

While the thesis argued that a lack of dialogue or collaboration between the creators of web archives, and end users (or even potential end users) has had some effect on engagement with web archives for research, there are various other implications. Certainly, what is evident throughout the thesis is the fact that the circumstances (legal, ethical, curatorial,

financial, technical, temporal, geographical, 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 (Winters, 2020a; Winters, 2019; Hockx-Yu, 2014; Gooding et al., 2021; Vlassenroot et al, 2019; Graham, 2017; Ogden & Maemura, 2017; Ogden, 2021; Brügger, 2021c; Ben-David, 2019; Ben-David, 2021). Therefore, the thesis argues that in order to understand the challenges for scholarly engagement with web archives, it is equally necessary to understand the challenges for web archive creators and how these challenges overlap and intersect across communities of practice within web archive research.

7.2 Revisiting the Research Questions and Answers

In the next sections, the research questions and answers are revisited and organised as follows:

- **7.2.1 Main causes for the loss of digital heritage:**
    - **RQ1**: What are the main causes for the loss of digital heritage? and how does this relate to Ireland?
- **7.2.2 Availability and accessibility of web archives based on the island of Ireland for conducting Irish based research:**
    - **RQ3**: How available and accessible are web archives based on the island of Ireland for conducting Irish based research?
- **7.2.3 Challenges and prospective solutions for participation in web archive research:**
    - **RQ2**: What are the main challenges for participation in web archive research? and how does this relate to Ireland?
    - **RQ4**: What is the current level of awareness of, and engagement and non-engagement with web archives in Irish academic institutions?
    - **RQ5**: How can we improve the conditions for conducting web archive research, and how does this relate to Ireland?

7.2.1 Main causes for the loss of digital heritage

Within a few years of the web becoming established as a new medium for publishing and sharing information, national libraries and cultural heritage organisations became concerned about the ephemeral nature of the web, and instigated preservational strategies for the capture and preservation of digital heritage on the web through web archiving.

Chapter 2.0 presented an overview of how these concerns were further substantiated by studies which examine link rot, reference rot, and web content change over time. There are several reasons put forward as to why web content moves, changes, or gets deleted, including software and system upgrading, changes in filing systems, the re-arrangement of web content, the relocation of servers, a lack of funding or interest to maintain websites, and simply a lack of foresight by web publishers. Chapter 2.0 also demonstrated how concerns for the loss of digital heritage on the web stemmed from wider concerns about the appraisal, storage and long-term preservation of electronic information, multimedia and born digital materials in general. Moreover, these wider concerns have been around since before the web was invented, with the web just becoming another media carrier to worry about.

As pointed out in chapter 2.0, UNESCO (2003) posits that "the disappearance of heritage in whatever form constitutes an impoverishment of the heritage of all nations", and digital heritage should not be an exception. Some of the factors which contribute to the loss of digital heritage to posterity include technological obsolescence of hardware and software, media deterioration, availability of resources, and inadequate legislation (UNESCO, 2003; Waters & Garrett, 1996; Besser, 2000). The loss of digital heritage has often gone unnoticed by societies and nations because "Attitudinal change has fallen behind technological change" and consequently, the economic, social, intellectual, and cultural value or potential value of the heritage is not realised (UNESCO, 2003). For Lyman (2002), societies have lost important parts of their cultural heritage in the past because it was not archived or preserved due to cultural, technical, economic, and legal problems. The cultural problem is due to the inability of past generations to recognise its importance and historic value, while the technical problem is due to a lack of foresight and technical ingenuity to ensure continuity for preservation, storage, and maintenance (Lyman, 2002). Lyman (2002) points out how the economic problem stems from the failure to find a business model to support the archiving of new media formats, while the legal problem stems from the failure to create legislation which protects copyright while at the same time allowing for archival preservation. These problems equally apply to the loss of digital heritage on the web. Although, it could be argued that the web archiving community has come a long way in providing solutions to the technical problem. Nonetheless, as internet and web technologies keep evolving, the capture tools will always be trying to catch up (Truman, 2016).

Chapter 2.0 also demonstrated how the evolving nature of publishing over the past 50 years became problematic for legal deposit legislation which was fundamentally print-centric. For hundreds of years the concept of legal deposit served as a system to compile and preserve

a collection of a country's publications outputs, thus providing a significant contribution to national cultural heritage. As a result, several countries began to amend their copyright and legal deposit legislation from the 1990s to accommodate the deposit of non-print materials and media formats (e.g., microfilm, CD-ROM, DVD etc.) and for born digital materials, inclusive of the web archiving of a country's national web domain, as a matter of routine. On the other hand, for many countries legal deposit legislation is still outdated in line with emerging publishing technologies and the advances in internet and web technologies. For example, the ROI has trailed behind Canada, New Zealand, and much of Europe (Conul, 2012, p. 14). Also, in relation to the ROI, chapter 2.0 highlighted how there have been continual warnings by the National Archives of Ireland (NAI) to the ROI government, since at least 1997, regarding the loss of digital heritage due to the lack of a "comprehensive formal records, management policy for State" and the "Loss of electronic records and archives or access to them, due to degeneration of storage media and/or redundancy of operating systems" (Reports of the Director of the National Archives of Ireland, 2014-2020). Regrettably, over twenty years since the problem was identified, the Irish government has still not come to terms with the preservation of electronic records, nor does it seem to have a formal policy for record keeping in any electronic format. Chapter 2.0 also demonstrated how content on the Irish government website(s) has changed and disappeared over the past decades, while chapter 4.0 established how Irish government department websites have been particularly vulnerable to link rot, and changes in website content.

Chapter 5.0 underscored one of the major causes for the loss of Irish digital heritage to posterity: the failure of successive ROI governments to negotiate copyright and legal deposit legislation in line with advances in publishing and communications technologies, exacerbated by the current deficiencies of ROI copyright and legal deposit legislation to include the routine web archiving of the Irish national domain as part of a national legal deposit scheme. In terms of the web space of NI, chapter 5.0 illustrated how the UK legal deposit legislation was reformed in 2003, to allow for a selective web archiving initiative (undertaken by the UK Web Archiving Consortium) which also incorporated the capture and preservation of websites from the NI web space. Moreover, the legislation was updated again in 2013 to allow for an annual web crawl of the UK web estate (undertaken by the UK Web Archive), inclusive of the NI web space. It also demonstrated how the PRONI Web Archive commenced a selective web archiving initiative in 2010 to capture and preserve websites of NI government departments, local councils, public sector organisations and websites which have social, cultural, political, religious, or economic significance for the preservation of NI heritage. However, prior to 2013, the UK/NI web space was not

systematically captured as part of legal deposit, and therefore much of the earlier NI webspace will have disappeared or changed drastically (Jackson, 2015a). To salvage some of the UK web estate prior to 2013, the Joint Information Systems Committee (JISC) acquired a dataset from the Internet Archive which included all .uk websites in their web collection that were crawled from 1996-2013 (UK Web Archive, n.d., JISC UK Web Domain Dataset). The JISC UK Web Domain Dataset is available for use through the UK Web Archive website and listed in the British Library Shared Research Repository.

7.2.2 Availability and accessibility of web archives based on the island of Ireland for conducting Irish based research

Chapter 5.0 examined the availability and accessibility of web archiving initiatives based on the island of Ireland, and their usefulness for conducting Irish based research. In doing so, it offered insights which may be useful when it comes to assessing support and incentive mechanisms for scholarly researchers using web archives and other types of end users. While the section acknowledged that Irish web heritage can be found in various international web archives, the focus was on web archiving initiatives which have a specific mandate to capture a wide range of Irish web heritage as part of their collection development strategies. Therefore, the focus was on the PRONI Web Archive, the NLI Web Archive, and the UK Web Archive, which is accessible onsite in the Library of Trinity College Dublin (TCD). The section observed the efforts of these initiatives for the collection and preservation of digital heritage from the web spaces of NI and ROI, and offered an overview of their historical backgrounds, inclusive of how copyright and legal deposit has influenced their collecting activities.

In the case of the digital heritage of NI, the findings suggest that, while there are resource and legislative limitations, there are nonetheless concrete efforts being made to provide a balanced approach towards the collection and preservation of the NI web space. First, the UK Web Archive captures and preserves websites from the NI web space, through a selective collection approach and through an annual domain crawl of the NI web space as part of legal deposit, which is accessible onsite in a UK legal deposit library, inclusive of the Library of TCD in Dublin. Second, NI digital heritage is preserved through a two-fold approach by PRONI to provide a publicly accessible selective web archive collection, through (i) the collection of websites of government, public bodies etc., with notifications of the intent to collect, and provisions of a takedown policy, and (ii) a permissions-based approach for privately funded websites. And third, the NI web space is preserved through a collaborative effort by the PRONI Web Archive and the UK Web archive for the development of accessible curated

collections. While there is a wide range of topics within the collections of the UK Web Archive and the PRONI Web Archive which would be useful for conducting Irish based research, access to the collections differ. PRONI Web Archive is open access, and the UK Web Archive is a mix of both open access and onsite access. However, as discussed, onsite access presents challenges for researchers due to the restrictive nature of the access protocols in the current UK legal deposit legislation, which is outdated in line with advances in publishing and communications technologies, and current trends in digital user expectations and information seeking behaviours (Gooding et al., 2019).

As regards the ROI, chapter 5.0 established how the National Library of Ireland (NLI) began a small-scale selective web archiving initiative in 2011, to include a wide range of topics which would be useful for conducting Irish based research. However, the chapter also highlighted how the NLI conducted two domain crawls in 2007 and 2017 which are currently inaccessible to researchers or the public due to legislative matters. While the NLI is a legal deposit library, digital legal deposit legislation was not enacted in Ireland at the time the domain crawls were conducted. Moreover, while digital legal deposit legislation came into force in December 2019 through the Copyright and Other Intellectual Property Law Provisions Act 2019 (hereafter, COIPLPA, 2019), it did not include a clause for crawling the Irish national web domain. However, COIPLPA (2019) does contain a clause to "bring forward a report on the feasibility of establishing a digital legal deposit scheme to serve as a web archive for .ie domain contents and advise on steps taken towards that goal" within twelve months of the Act coming into force in December 2019. As of October 2022, a feasibility report has yet to be produced.

Chapter 5.0 highlighted how the establishment of a ROI national domain web archive is a necessary component for the preservation of Irish national digital heritage and examined some of the political debates regarding the inclusion of the archiving of the Irish national web domain as part of legal deposit legislation in line with other countries. It outlined that, while there are traces of the ROI web estate in other web archives, these are very shallow. Therefore, it will be impossible to retrospectively recreate the ROI web space. As it stands, the ROI is already "impoverished" (UNESCO, 2003) due to mass losses of digital heritage on the web for the decades of the 1990s, 2000s, and 2010s. It now looks like this will continue well into the 2020s, before the necessary measures are put in place for the collection and preservation of the web space of the twenty-six counties of the ROI in line with the collection and preservation of the web space of the six counties of NI. Therefore, it was stressed that immediate action is required for an emergency change in ROI legislation to allow for the collection and preservation of the ROI web estate in the interim, while a feasibility report

continues to be undertaken to advise on the necessary requirements to update the legislation, and to establish a national web domain archive through “a process of negotiation among interested parties” (Lyman, 2002). Moreover, as demonstrated, negotiations should be inclusive of representatives from the education and teaching sectors, end users who use web archives for wide range of purposes, information professionals who have experienced the transition from small-scale selective web archiving to large-scale domain web archiving, and information professionals who are experienced in working with Irish based information ecosystems.

Chapter 5.0 further emphasised the need to assess the demarcation of the Irish national web domain, as using the .ie ccTLD is not an adequate marker for the representation of Irish digital heritage on the web. Finally, the section underscored how born digital content is more fragile than print material, and publishing and communications technologies are constantly changing. Thus legal deposit legislation needs to be reviewed on a regular basis in order to keep up with the changes in technology and current trends in digital user expectations and information seeking behaviours (Gooding et al., 2019). The chapter further noted how the formation of a Copyright Council of Ireland, as suggested in the Modernising Copyright report (2013), could be tasked with monitoring legal deposit legislation in line with the fragility of born digital heritage and the technological advances in publishing and communication technologies.

7.2.3 Challenges for participation in web archive research and prospective solutions, and how this relates to Ireland

The thesis identified 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. These are equally applicable for Irish based web archiving communities, as well as Irish based researchers who engage with, or might potentially engage with, the archived web for research purposes. This section offers an overview of the findings in terms of the main challenges for participation in web archive research, and potential solutions for improving the conditions for conducting web archive research. Within this, the findings are also presented regarding the current levels of awareness and engagement and non-engagement with web archives in Irish academic institutions.

Chapter 3.0 demonstrated how institutional web archiving is a complex process that necessitates much decision making. Decisions must be made on the appraisal and selection of content to be captured; the technology to use for capturing, storage, preservation, and

replay/playback; as well as how to make the collected data accessible for use, and indeed, how flexible this access might be. Furthermore, such decisions may be influenced by social, cultural, and political circumstances; legislations on copyright and legal deposit or lack thereof; and the availability of resources in terms of finance, labour, technology, and organisational infrastructures (Ogden, 2021; Dougherty, 2007; Ben-David, 2019; Ben-David, 2021; Hockx-Yu, 2014; Winters, 2020a; Winters, 2019; Vlassenroot et al., 2019; Brügger, 2021c; Maemura, 2022). Web archiving is further complicated by "ever-evolving" internet, web, and software technologies, thus, such technologies "will always be ahead of the capture tools" (Truman, 2016). Therefore, the web archiving life cycle of tools will keep changing too.

Chapter 3.0 further examined how search and retrieval capabilities have presented challenges, and while the web archiving community has worked on improving its search capabilities through complementing traditional URL search with metadata and full-text search, they have encountered considerable challenges along the way. Metadata search entails a search through various types of attributes such as a subject category, a description, a language, or file format. However, the "manual creation" of descriptive metadata for selective curated collections is resource intensive, making it "a non-viable option" for large-scale web archives. Therefore, metadata needs to be created automatically for large-scale collections (Costa, 2021). 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). 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 (Costa, 2021; Holzmann & Nejdl, 2021; Winters & Prescott, 2019; Jackson et al., 2016; Nielsen, 2016). Because web archive search also includes a temporal dimension, algorithms that were developed for ranking search results from the live web will not provide satisfactory results. Therefore, for Jackson et al. (2016b) if a ranking model is used it must be made completely transparent so scholars can interpret the results accordingly.

In essence, due to the ever-changing landscape in internet, web, and software technologies, web archiving initiatives depend on continual research on crawler-based archiving and techniques for improving crawler efficiency to enable better data quality assurances, as well as techniques and software developments for search and retrieval, replay/playback, digital preservation, software archaeology, IT integrations, and more (Denev et al., 2009; Spaniol et al., 2009; Denev et al., 2011; Xie et al., 2013; Bingham, 2014; Mourão & Gomes, 2021; Newing & Clegg, 2021; Samar et al. 2017; Jackson, 2022a; Jackson, 2022b; UK Web Archive,

2018; Day, 2006; Alberts et al., 2017; Jansma, 2020; Beis et al., 2019). Moreover, web archiving initiatives also have challenges for the collection of web content, and the provision of access to web content due to legalities such as outdated copyright and legal deposit legislation and ethical and privacy concerns (Graham, 2017; Hock-Yu, 2014; Ryan et al., 2022). Nonetheless, continual efforts are being made by heritage organisations and web archive curators to capture what they can, as best they can (Laursen & Møldrup-Dalum, 2017).

Chapter 3.0 also demonstrated several challenges with permissions-based selective collections. Not all websites provide contact details and even if a contact is found there is no guarantee that a website owner will respond (Ryan et al., 2022; Bingham & Byrne, 2021). Pennock (2013) and Brown (2006) point to the weaknesses of selective web archiving due to selector bias (albeit it unintentional or unacknowledged). Brown (2006) notes how the sheer size and depth of the web makes it difficult for manual selectors to stay abreast of evolving sources, and subject knowledge. Moreover, chapter 2.0 pointed out how the concepts of in-groups and out-groups are acknowledged as a phenomenon of human behaviour which exhibit in-group favouritism, and discrimination towards out-groups (Tajfel 1970; Tajfel 1971; Tajfel et al., 1974). This will also influence what is included or excluded as part of a selective thematic collection. Chapter 2.0 argued that, to combat these issues, legal deposit libraries across Europe opt to conduct both selective and domain-wide web archiving as a more balanced, representative and inclusive approach towards the capture of national digital heritage on the web.

Chapter 3.0 then examined reasons for the lack of scholarly engagement with web archives, and the challenges for end user/researchers, and put forward several reasons for the lack of scholarly 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; Winters, 2017; Costea, 2018). It can also be argued that a lack of dialogue or collaboration between the creators of web archives, and end users (or even potential end users) has had some effect on engagement with web archives for research purposes as, initially, web archiving strategies tended not to prioritise how web archives would or might be used (Dougherty et al., 2010; Hockx-Yu, 2014; Schroeder & Brügger, 2017; Gooding et al., 2021). Thus, Truman (2016) stresses the need for more communication and collaboration between those who curate, create and steward web archives and those who use (or might use) a web archive for purposeful research.

Challenges also arise, due to the characteristics of an archived website or web page which may not be a complete surrogate of what was once on the live web, rather, it is a version (Brügger, 2010). Deficiencies in the archived artefacts may occur because of the temporal dimensions such as the time it takes to capture, and the possibility of content updates during capture. Deficiencies may also occur due to technical issues such as glitches during the archiving process that involve robots.txt or limitations with the archiving software/hardware to keep up with the constant change and upgrade of web media file types and the evolving nature of dynamic content (Brügger, 2010; Meyer et al., 2011; Pennock, 2013; Maemura, 2018; Bingham & Byrne, 2021). In addition, in order to preserve a website or web page in its entire capacity to produce meaning, it should be inclusive of links to external (hyperlink) information, and quite often this is not achieved due to selection criteria, acquisition policies, technical glitches, financial constraints, or legislative and copyright restrictions (Besser, 2000; Milligan, 2019; Hockx-Yu, 2014). Finally, the collected web content may undergo technical processes during collection, preservation and the provision of access through replay or playback (Brügger 2016, 2018; Schneider et al., 2009). This is why Brügger (2019; 2018; 2016) describes archived web content as reborn digital media, which is clearly distinct from other types of archived media such as film, television, photographs, and newspapers. Therefore, this implies that the use of archived web content for scholarly purposes has ongoing pedagogical challenges.

Other commentators note challenges due to the variances between searching on the live web, and searching in a web archive (Costa, 2021; Holzmann & Nejdl, 2021; Winters & Prescott, 2019; Jackson et al., 2016b; Nielsen, 2016). The findings through web archive search techniques also tend to present multiple copies of content captured during different crawls, so they have a temporal dimension, which manifests more challenges. Both Brügger (2016) and Schafer (2019) suggest that web archives present challenges due to the "absence" of a traditional style catalogue or registry as an entry point. Costea (2018) identifies a need for improvements to web archives in the areas of discoverability options, data selection, data management, and access to more comprehensive documentation and metadata. Challenges for researchers/users also arise 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 (Truman, 2016). Researchers wishing to take a more qualitative approach towards using the archived web also have challenges due to a lack of research methods and theoretical paradigms for the use of the archived web (Millward, 2015; see Table 4.15). Other challenges relate to the fact

that some large-scale web archives, such as the Internet Archive's Wayback Machine, may lack depth and are deemed as too broad to meet the needs of specific research which often requires precise datasets (Schneider et al., 2009; Dougherty & van den Heuvel, 2009). Therefore, researchers often turn to developing their own web archive collections for their needs (see for example, Foot & Schneider, 2006; Engholm, 2000). However, such collections are often narrow in scope and may never be useful for anything other than the study for which they were created (Dougherty & van den Heuvel, 2009).

Legislation on copyright and legal deposit also presents challenges for researchers to utilise web archives. Using the UK Web Archive legal deposit collections as an example, scholars discuss the challenges in using legal deposit collections which are only accessible on a library terminal in a designated reading room. Such challenges include the locked down nature of the library terminal whereby researchers cannot view the source code or copy the URL from the browser which causes problems for citation (Winters, 2020a; Milligan, 2015). Users are not allowed to copy and paste text which totally disrupts the affordances that are used by researchers worldwide, when they use the live web as a source for research (Milligan, 2015). Also, users can not take photographs or screenshots of the screen, rather they must pay for a printout of an archived web page, which is ironic, as researchers are allowed to use cameras to take photographs of historical documents in most archival environments (Milligan, 2015). Furthermore, no two people can view the same instance of an archived web page simultaneously which inhibits collaborative research as well as the use of the resource for teaching in the context of classroom group projects (Winters, 2020). Such challenges are manifested due to the restrictive nature of the UK legal deposit legislation as laid out in The Legal Deposit Libraries (Non-Print Works) Regulations 2013 (NPLD). Gooding et al. (2019) also discuss the challenges with the NPLD access protocols and highlight how the NPLD regulations make no allowance for text or data mining, and how this presents a barrier for innovative research. Furthermore, Gooding et al. (2021) suggest that the user was neglected as a stakeholder when it came to drafting the legislation for NPLD access protocols, which is fundamentally print-centric. Moreover, they insist that because the NPLD ethos is print-centric, it fails to consider the user in line with digital user expectations, and current trends in information seeking behaviours (Gooding et al., 2021). Therefore, when it comes to evaluating resources like legal deposit collections, in particular the use of collections with restrictions, it needs to be clearly examined in relation to the rapidity in which technology changes the landscape for end users.

There are other implications regarding the use of web archives with access restrictions. Maurer (2022) and Healy et al. (2022) note how the provision of onsite 'only' access to web

archive collections in a designated building makes web archives geographically and socio-economically inaccessible for many researchers. Furthermore, Truter (2021) highlights the challenges for end user researchers in terms of the access and use of archived web content 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). This manifests challenges for not only the use of the data, but also affects how and if the data can be made shareable and reusable (Truter, 2021) and runs counter to the requirement of open science which is being stipulated by a growing number of research institutions and funding agencies (Winters, 2020a).

Chapter 3.0 discussed how challenges for researchers arise due to ethical, sociotechnical, and political circumstances. Maemura (2018) 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. Ogden et al. (2022) suggest that researchers need to be vigilant using web archives when researching socially vulnerable communities, and Mackinnon (2021) warns researchers of the ethical implications when it comes to the study of websites of "young people of the past" and their right to be forgotten. Ogden and Maemura (2021) examine how the sociotechnical, organisational, and resource constraints "under which most web archiving programmes operate" need to be understood by researchers, and suggest that researchers need to become familiar with the "specific limits and constraints, legal governance frameworks, collection mandates, as well as configurations (i.e. of sub-collections) and terminology used for specific collections." In terms of political circumstances, Ben-David (2019) discusses the challenges for studying web histories of countries that do not have a ccTLD, such as Kosovo, which was denied the allocation of a ccTLD as it was not recognised as a sovereign state by the United Nations, due to a Russian veto.

Chapter 3.0 highlighted how researchers may also be more interested in using big data methods such as topic modelling or network analysis on a web sphere of websites (WARC files) from a specific web archive collection (e.g., Geocities) or to do a longitudinal study across multiple legal deposit annual web domain collections (see Milligan, 2019; Brügger et al., 2017; Brügger et al., 2019). However, Maurer (2022) points out that organising large volumes of WARC files for research is difficult for both web archiving initiatives, and end user researchers. Reasons for this are varied and may be "due to a mix of curatorial, technical, legal, economic and organisational constraints" (Brügger, 2021c). Brügger (2021c) further stresses the need for solid research infrastructures between the web archives with the data, and the research teams wishing to use the data, to help overcome some of the

legal, ethical, and technical challenges for both communities. This will require funding, and a cultural shift placing the creator and user as partners in the full web archiving lifecycle.

Chapter 4.0 posited that one should also consider how some of the challenges mentioned above overlap between creators and users. For example, both creators and users have challenges in the areas of search and retrievability, users find it difficult to search large-scale web archives, while creators find it difficult to provide search mechanisms and algorithms for large-scale collections that will satisfy a diversity of users. Moreover, both creators and users have challenges with legal issues such as copyright and legal deposit, users have challenges accessing content, while creators have challenges for the collection of web content, as well as the provision of access, and how restrictive this access might be.

Through an online survey, chapter 4.0 focused 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. The chapter explored the skills, tools, and knowledge ecologies in web archive research, and examined the challenges for participation in web archive research, and the overlaps and intersections of such challenges across communities of practice. In doing so, it organised the participants into two thematic representations of 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. The findings presented a wide range of challenges experienced by both communities. Participants in a library, archive, or web archive environment experienced challenges such as: inconsistencies and incompleteness; legalities for acquisition/access; challenges with learning new skills; producing documentation/metadata; volume of data; institutional challenges; and technical challenges. While participants who identified as being a scholar, academic, lecturer, student, or working in an IT/web design environment experienced challenges such as: inconsistencies and incompleteness; legalities on access, use, and storage; challenges with learning new skills; lack of documentation/metadata; volume of data for research; challenges in an IT/business/admin. environment; performance related issues; and challenges for research methods and approaches.

The findings presented several commonalities between participants from both communities. 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 findings 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 addition, the findings illustrated how multiple challenges 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 PDFs 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. It was noted how the provisions of fully comprehensive metadata are problematic when dealing with high volumes of crawled data, as they are 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.

Chapter 4.0 also demonstrated how participants from both communities experience challenges when citing content from a web archive and point to a lack of guidelines, standards, or best practices, for citing archived web materials, and for some participants it is simply not easy to cite sources from a web archive. Participants also noted challenges for

citing materials from a legal deposit archive, or archives with restrictive access. Thus, this becomes problematic for the transparency of the research methods being used. Participants also described challenges for citing datasets of archived web content due to a lack of guidelines and standards for citing datasets, and some participants indicated that it is not easy to cite datasets in general. Other concerns relate to the data/content reliability of a dataset in terms of its page capture/completeness, and preservation reliability is also mentioned by one respondent. The citation 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.

In terms of tools and methods, the findings suggested how 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 findings specified how 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. Survey participants from a scholarly or academic environment engaged with a diversity of tools and methods, and the research question or methodology often influenced which tools and methods were 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 findings from chapter 4.0 demonstrated how challenges in learning new skills are also experienced by respondents from both communities, and therefore it was suggested how

both communities would benefit from the provision of collaborative communal training across the full range of activities in the web archiving lifecycle. The chapter also offered 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. It was proposed 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. It was further recommended 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. Moreover, the chapter found that the challenges experienced by the survey participants did not diminish with increasing experience and highlighted the need for training across all levels of experience. It was indicated 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. Chapter 4.0 also pointed out how training and education are directed towards the latest communications technologies, negating the need for training and education in the long-term preservation of information created by these new technologies.

Chapter 4.0 discussed how legalities, such as legal deposit, copyright, and GDPR present other challenges for both the web archiving and researcher/user communities. 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 run counter to current trends by funders, who are increasingly stipulating for open access and open science frameworks for research and data outputs. The chapter suggested that further discussion and collaboration is required to foster developments in the areas of the application of research data management practices 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, chapter 4.0 provided positive acknowledgements which reinforce the need and the value of collaborations across communities of practice and highlighted how collaborations between web archive creators and users/researchers can benefit both communities in addressing some of the challenges mentioned above. However, it was also acknowledged 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, 2021c). 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, legal professionals or lawyers. This will be equally important for the successful establishment of an Irish web domain archive, from creation to end use.

Chapter 6.0 offered some insights into the challenges and solutions for web archive research in an Irish context. Through an online survey of lecturers, researchers, and students in Irish academic institutions, the section set out to provide some insight into the awareness of, and engagement with, web archives in Irish third-level academic institutions, in a bid to gain a better understanding of how and why archived web content is used or not used for research

in Ireland. The section was also exploratory in terms of assessing some of the opportunities and challenges for using web archives, and considerations for how to best facilitate their use, going forward. Most prominently, the findings demonstrated a limited awareness of the existence of web archives in Irish academic institutions, and that creating awareness increases the probable likelihood for an increase in researcher engagement. However, the findings suggested that promoting awareness of the existence of web archives by itself may not be sufficient to impact engagement. For an unfamiliar audience, efforts are also needed to demonstrate the importance of archiving the web, the value of web archives for research, and more effort for awareness on how to use web archives for research. The findings also indicated that web archives will become more important for research as time goes on.

The survey findings also present several indicators on the challenges that scholars based in Ireland perceive for the future use of web archives and archived web content. How to use web archives and archived web content was presented as a challenge from several outlooks such as search and navigation, handling large volumes of data, citation practices for using archived web content, and research models for using web archives as a non-established source. Of interest are the different outlooks on the use of large-scale analysis, and the need for training in big data analysis for Humanities, while there are also concerns that big data analysis does not account for a full understanding of the context of the data. Rather, this might be better achieved with a qualitative approach. This implies that there is a need to consider research models that consider both qualitative and quantitative methods as standalone practices, or a mixture of both as a combined approach to include web archives as a resource for research in Ireland. The completeness of the data was also mentioned in terms of capture frequencies, as well as challenges with the representativeness of the data in a web archive, in relation to what is presented (or not presented) on the web and what ends up in a web archive. There is also the case that data in a web archive is simply not relevant for a particular research discipline.

On a bright note, the findings also show that there is already a small community of web archive users in Irish academic institutions, aged from 18 to 65 years, and at different levels of education and academia. The findings further indicated that user respondents utilise web archives and archived web content for coursework purposes, for professional publication and historical research purposes, for teaching purposes, for qualitative and quantitative research purposes and for access to materials no longer available on the live web. What is also surprising is that users come from a diverse range of research fields, which reflects the need for both multidisciplinary and interdisciplinary deliberation to consider the challenges, and potential solutions, for developing research models and paradigms for the use of web

archives for Irish based research that are fit for purpose in a broad spectrum of research fields.

Certainly, the user responses in this study offer some valuable insights on the opportunities for the use of web archives for Irish based research, and there is reason to believe that this community will grow over the next few years, as more academics become aware of web archives as resources for research. However, increases in web archive engagement will also depend on the promotion of awareness of the value of web archives, and demonstrations of use cases in academia as well as the public sphere. Formulating an Irish based multidisciplinary/interdisciplinary research network to comprise of current scholarly users and potential users, web archivists, information professionals and web design professionals would be of great benefit here. It would assist in addressing potential solutions for developing research models and paradigms for the use of web archives for Irish based research across a broad spectrum of research fields; and enable discussions to develop frameworks to provide course modules for students in the use of web archives for research, and training courses for educators on how to incorporate web archived content as part of their teaching materials and methods.

To end here, the survey findings indicated that awareness of the NLI domain archive is quite poor, and thus, will warrant a strategy for promotion as a research resource, when it eventually becomes accessible. In this regard it will be essential for the NLI to be afforded the capacity to collaborate with users and promote the resource to potential users, and the capacity to build solid research infrastructures between the NLI web archive and the research teams seeking to use the data. This will require funding, and a cultural shift placing the creator and user as partners in the full web archiving lifecycle. In addition, access to an Irish domain web archive onsite in the NLI reading room 'only' will present geographical and socio-economic barriers for some researchers. Therefore, in terms of the establishment of an Irish domain web archive, the obvious solution to the access problem would be to make it open access using an 'Opt-Out' strategy. However, this is probably unlikely for all types of web content. Therefore, for content that requires restrictions, such as content behind paywalls, there will be a need to consider how access can be provided in more than one geographic location, perhaps in conjunction with other legal deposit libraries across Ireland. Moreover, access provisions should be made for researchers and users who are not affiliated to an academic institution. In the long-term, access should be provided in public libraries across Ireland, and this would ensure that users are not disadvantaged based on geographic location or socio-economic circumstances.

It must also be emphasised that certain categories of websites should be open access by default, including:

(i) websites belonging to the Irish government, its departments, and its subsidiary agencies, as well as local government and councils,
(ii) websites belonging to public bodies, quangos, civic agencies, and political parties who receive government funding in any form,
(iii) websites belonging to owners or organisations who have received funding from the Irish government or any of its subsidiary agencies, and this should be stipulated as part of any funding agreement, and
(iv) websites which have a variety of Creative Commons licences could also be considered for inclusion for open access.

7.3 Final Thoughts and Future Work

While this thesis sought to examine web archive research in line with international developments and how this relates to Ireland, it is by no means complete, as there will always be a need to continually examine the skills, tools, and knowledge ecologies within web archive research as long as internet web and software technologies keep advancing and changing. Rather, it is hoped that this thesis would serve as a starting point for fostering open dialogues across Ireland on the necessity for long-term preservation strategies for electronic information, multimedia, and born digital materials in general, with the web just being another media carrier to worry about. In laying the groundwork, the thesis therefore contributes to the current debates regarding the necessity for the implementation of legal deposit legislation which realistically reflects the fragility of born digital heritage and the technological advances in publishing and communication technologies.

The thesis also provides a starting point in addressing some of the challenges regarding digital and web historiography and how this relates to Irish based research. As new methodologies are born out of necessity to deal with the advances of the internet, web and software technologies and the continual evolution of digital media, older methodologies will be doomed due to software incompatibilities or obsolescence and outdated digital media formats. This is not something new. Archivists, librarians, and information professionals have been discussing it for years. The big question here now is how this affects the use of digital materials for academic research, whether they are digitised, born digital on the live web or reborn digital in a web archive and how can we ensure such digital materials remain accessible, allowing for research reproducibility in the future.

BIBLIOGRAPHY

The bibliography is organised in four parts: a list of **Primary Sources** and **References**, a list of web archiving **Providers & Services**, and **Software, Tools & Methods** that are mentioned throughout the thesis. The full Bibliography is also available in the Zotero web library for the doctoral project.[57] I have tried to ensure that the URLs provided in the Bibliography and footnotes are (i) captured in a web archive close to the time of access on the live web or (ii) saved in a web archive close to the time of access on the live web. In case of future link rot, I have documented which web 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 doctoral project files, available in Open Science Framework (https://osf.io/t42va/).

Primary Sources

Acts, Bills, Amendments, Directives, Statutes

European Communities. Council Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society. European Parliament, 167 OJ L (2001). Retrieved 2022-10-31, from http://data.europa.eu/eli/dir/2001/29/oj/eng [URL Memento: Archive.today] EUR-Lex Doc ID: 32001L0029

European Economic Community. Council Directive 91/250/EEC of 14 May 1991 on the legal protection of computer programs. European Parliament, 122 OJ (1991). Retrieved 2022-09-28, from https://op.europa.eu/en/publication-detail/-/publication/92d68447-ea9a-4554-9540-de517984c310/language-en. [URL Memento: Wayback Machine]

European Economic Community. Council Directive 93/98/EEC of 29 October 1993 harmonizing the term of protection of copyright and certain related rights. European Parliament, 290 OJ L (1993). Retrieved 2022-09-28, from http://data.europa.eu/eli/dir/1993/98/oj/eng. [URL Memento: Wayback Machine]

Ireland. Copyright Act, 1963. No. 10/1963, Government of Ireland, Oireachtas (1963). Commencement Date: 01 October 1964. Retrieved 2022-02-06, from https://www.irishstatutebook.ie/eli/1963/act/10/enacted/en/html. [URL Memento: Wayback Machine]

[57] Healy, S. (2022). Zotero Groups - The Future(s) of Web Archive Research Across Ireland. Zotero, https://www.zotero.org/groups/4712321/the_futures_of_web_archive_research_in_ireland_s.c._healy.

Ireland. Eighth Amendment of the Constitution Act, 1983. No. C8/1983, Government of Ireland, Oireachtas (1983), Commencement Date: 07 October 1983. Retrieved 2021-09-23, from https://www.irishstatutebook.ie/eli/1983/ca/8/enacted/en/index.html. [URL Memento: Wayback Machine]

Ireland. National Archives Act, 1986. No. 11/1986, Government of Ireland, Oireachtas (1986). Commencement Date: 01 June 1988. Retrieved 2022-03-22, from https://www.irishstatutebook.ie/eli/1986/act/11/enacted/en/html?q=National+Archives+Act [URL Memento: Wayback Machine]

Ireland. Dublin City University Act, 1989. No.15/1989, Government of Ireland, Oireachtas (1989). Commencement Date: 22 June 1989. Retrieved 2021-04-20, from https://www.irishstatutebook.ie/eli/1989/act/15/enacted/en/html. [URL Memento: Wayback Machine]

Ireland. University of Limerick Act, 1989. No.14/1989, Government of Ireland, Oireachtas (1989). Commencement: 22 June 1989. Retrieved 2021-04-20, from https://www.irishstatutebook.ie/eli/1989/act/14/enacted/en/html. [URL Memento: Wayback Machine]

Ireland. European Communities (Legal Protection of Computer Programs) Regulations, 1993. S.I. No. 26/1993 Government of Ireland, Oireachtas (1992). Commencement Date: 31 December 1992. Retrieved 2021-05-08, from https://www.irishstatutebook.ie/eli/1993/si/26/made/en/print. [URL Memento: Wayback Machine]

Ireland. European Communities (Term of Protection of Copyright) Regulations, 1995. S.I. No. 158/1995, Government of Ireland, Oireachtas (1992). Commencement Date: 01 July 1995. Retrieved 2021-10-16, from https://www.irishstatutebook.ie/eli/1995/si/158/made/en/print. [URL Memento: Wayback Machine]

Ireland. National Cultural Institutions Act, 1997, No.11/1997, Government of Ireland, Oireachtas (1997). Commencement Date: 02 June 1997. Retrieved 2022-01-19, from https://www.irishstatutebook.ie/eli/1997/act/11/enacted/en/html. [URL Memento: Wayback Machine]

Ireland. Universities Act, 1997. No.254/1997, Government of Ireland, Oireachtas (1997). Commencement Date: 16 June 1997. Retrieved 2021-08-01, from https://www.irishstatutebook.ie/eli/1997/act/24/enacted/en/index.html. [URL Memento: Wayback Machine]

Ireland. Copyright and Related Rights Act, 2000. No. 28/2000, Government of Ireland, Oireachtas (2000). Commencement Date: 01 January 2001. Retrieved 2019-03-27, from https://www.irishstatutebook.ie/eli/2000/act/28/enacted/en/html. [URL Memento: Wayback Machine]

Ireland. Copyright and Other Intellectual Property Law Provisions Bill 2018 (as initiated). No. 31/2018, Government of Ireland, Oireachtas (2018). As initiated: 09 March 2018. Retrieved 2022-01-20, from

https://data.oireachtas.ie/ie/oireachtas/bill/2018/31/eng/initiated/b3118d.pdf [URL Memento: Wayback Machine]

Ireland. Thirty-sixth Amendment of the Constitution Act 2018. Act C36/2018 Government of Ireland, Oireachtas (2018), Commencement Date: 18 September 2018. Retrieved 2021-02-05, from https://www.irishstatutebook.ie/eli/2018/ca/36/enacted/en/html. [URL Memento: Wayback Machine]

Ireland. Copyright and Other Intellectual Property Law Provisions Act 2019. No. 19/2019, Government of Ireland, Oireachtas (2019). Commencement Date: 02 December 2019 Retrieved 2021-10-26, from https://www.irishstatutebook.ie/eli/2019/act/19/enacted/en/index.html [URL Memento: Wayback Machine]

Irish Free State. Industrial and Commercial Property (Protection) Act, 1927. No. 16/1927 Irish Free State, Oireachtas (1927). Commencement Date: 01 October 1927. Retrieved 2021-04-25, from https://www.irishstatutebook.ie/eli/1927/act/16/enacted/en/html. [URL Memento: Wayback Machine]

Seanad Éireann. Copyright and Other Intellectual Property Law Provisions Bill 2018—Committee Stage—Amendments. No. 31b/2018, 03 October 2018, Houses of the Oireachtas, Ireland (2018). Retrieved 2022-01-20, from https://data.oireachtas.ie/ie/oireachtas/bill/2018/31/seanad/3/amendment/numberedList/eng/b31b18d-scnl.pdf [URL Memento: Wayback Machine]

Seanad Éireann. Copyright and Other Intellectual Property Law Provisions Bill 2018—Report Amendments. No. 31b/2018, 15 May 2019, Houses of the Oireachtas, Ireland (2019). Retrieved 2021-04-19, from https://data.oireachtas.ie/ie/oireachtas/bill/2018/31/seanad/4/amendment/numberedList/eng/b31b18d-srnl.pdf [URL Memento: Wayback Machine]

United Kingdom. Act of Union (Ireland) 1800. Acts of the Old Irish Parliament 1800 c. 38 (Regnal. 40_Geo_3), UK Parliament (1801). Commencement Date: 01 January 1801. https://www.legislation.gov.uk/aip/Geo3/40/38/contents. [URL Memento: Wayback Machine]

United Kingdom. Public Records Act (Northern Ireland) 1923. Acts of the Northern Ireland Parliament 1923, Chapter 20, UK Parliament (1923). Commencement Date: 22 June 1923. Retrieved 2020-08-05, from https://www.legislation.gov.uk/apni/1923/20/contents. [URL Memento: Wayback Machine]

United Kingdom. Legal Deposit Libraries Act 2003. UK Public General Acts 2003 c. 28, UK Parliament (2003). Commencement Date: 01 February 2004. Retrieved 2022-01-08, from https://www.legislation.gov.uk/ukpga/2003/28/contents. [URL Memento: Wayback Machine]

United Kingdom. The Legal Deposit Libraries (Non-Print Works) Regulations 2013. UK Statutory Instruments 2013 No. 777, UK Parliament (2013). Commencement Date: 05 April 2013. Retrieved 2022-02-03, from

https://www.legislation.gov.uk/uksi/2013/777/contents/made. [URL Memento: Wayback Machine]

E-Mail

IIPC curators list. (2017, April 4). Call for Participation: Archives Unleashed 4.0: Web Archive Datathon [distribution list communication].

E-Zines

Sterne, J. (1995, February 27). IT's Monday One Hundred and Forty-Two—27 February 95. *IT's Monday*, 142. TechArchives, Ireland. TechArchives, Ireland - Repository

Newspapers

Crowe, C. (2012, June 30). Ruin of Public Record Office marked loss of great archive. *The Irish Times* [online] Retrieved from https://www.irishtimes.com/opinion/ruin-of-public-record-office-marked-loss-of-great-archive-1.1069843 [URL Memento: Wayback Machine]

Cunningham, M. (1997a, March 17). Beware the ideas of March The Government Web site (http://www.irlgov.ie). *The Irish Times*, (COMPUTIMES), [CITY EDITION], p. 10. Proquest ID: 310308514

Cunningham, M. (1997b, January 27). Brewster's millions. *The Irish Times*, (COMPUTIMES), [CITY EDITION], p. 18. ProQuest ID: 310187147

Cunningham, M. (1997b, January 27). Brewster's millions. *The Irish Times*, (COMPUTIMES) [online] Retrieved from https://web.archive.org/web/19990117002422/http://www.irish-times.com/irish-times/paper/1997/0127/cmp1.html. [Wayback Machine, timestamp: 1999-01-17 00:24:22; URL source: http://www.irish-times.com/irish-times/paper/1997/0127/cmp1.html]

Cunningham, M. (1995, January 2). 1994: Year of the Net Michael Cunningham chronicles some of the main computing stories during annus mulllmedius. *The Irish Times*, (COMPUTIMES), [CITY EDITION], p. 17. ProQuest ID: 309949150

Fagan, K. (2012, May 21). When the server becomes the master. *The Irish Times* [online] Retrieved 2022-10-31, from https://www.irishtimes.com/business/sectors/when-the-server-becomes-the-master-1.522745. [URL Memento: Wayback Machine]

Finn, C. (2022a, March 14). McDonald says deletion of statements on SF website not attempt to pivot position on Russia. *The Journal* [online].Retrieved 2022-03-15, from https://www.thejournal.ie/mcdonald-russia-sinn-fein-ukraine-5710841-Mar2022/. [URL Memento: Wayback Machine]

Finn, C. (2022b, March 15). Taoiseach says removal of certain press statements from SF website is 'kind of Orwellian'. *The Journal* [online]. Retrieved 2022-03-15 from https://www.thejournal.ie/sinn-fein-deleting-statements-taoiseach-5712160-Mar2022/. [URL Memento: Wayback Machine]

Gataveckaite, G. (2022, March 14). Sinn Féin deletes thousands of statements from its website due to 'outdated content'. *The Irish Independent* [online]. Retrieved 2022-03-18, from https://www.independent.ie/irish-news/politics/sinn-fein-deletes-thousands-of-statements-from-its-website-due-to-outdated-content-41443385.html. [URL Memento: Wayback Machine]

Horne, E. (2016, November 5). The great flood of Florence, 50 years on. *The Guardian*. Retrieved 2022-03-24, from https://www.theguardian.com/artanddesign/2016/nov/05/the-great-flood-of-florence-50-years-on. [URL Memento: Wayback Machine]

McDonald, H. (2008, April 2). Irish prime minister Ahern resigns amid financial controversy. *The Guardian* [online]. Retrieved 2021-10-23, from https://www.theguardian.com/world/2008/apr/02/ireland [URL Memento: Wayback Machine]

McGee, H. (2018, December 28). Move to begin in 2019 to release State papers after 20 years. *The Irish Times* [online]. Retrieved 2018-12-28, from https://www.irishtimes.com/news/politics/move-to-begin-in-2019-to-release-state-papers-after-20-years-1.3742288 [URL Memento: Wayback Machine]

O'Connell, H. (2022, March 13). Sinn Féin wipes years of media statements from website. *Sunday Independent* [online]. Retrieved 2022-03-13, from https://www.independent.ie/irish-news/news/sinn-fein-wipes-years-of-media-statements-from-website-41440873.html. [URL Memento: Wayback Machine]

Shenton, H. (2020, April 29). Digital black hole in our national memory. *The Irish Times* [online]. Retrieved 2020-05-14, from https://www.irishtimes.com/opinion/letters/digital-black-hole-in-our-national-memory-1.4240247 [URL Memento: Wayback Machine]

Ramesh, R., & Hern, A. (2013, November 13). Conservative party deletes archive of speeches from internet. *The Guardian* [online]. Retrieved 2020-06-13, from https://www.theguardian.com/politics/2013/nov/13/conservative-party-archive-speeches-internet. [URL Memento: Wayback Machine]

Taylor, C. (2017a, July 20). Ireland's digital content in danger of disappearing, specialist warns. *The Irish Times* [online]. Retrieved 2017-09-22, from https://www.irishtimes.com/business/technology/ireland-s-digital-content-in-danger-of-disappearing-specialist-warns-1.3157792. [URL Memento: Wayback Machine]

The Irish News. (2022, March 14). Sinn Féin removes thousands of media statements from its website. *The Irish News* [online]. Retrieved 2022-03-14, from https://www.irishnews.com/news/northernirelandnews/2022/03/14/news/sinn-fe-in-removes-thousands-of-media-statements-from-its-website-2613771/. [URL Memento: Wayback Machine]

Weiss, R. (2003, November 24). On the Web, Research Work Proves Ephemeral. *Washington Post* [online]. Retrieved 2018-04-29, from https://www.washingtonpost.com/archive/politics/2003/11/24/on-the-web-research-work-proves-ephemeral/959c882f-9ad0-4b36-88cd-fb7411db118d/ [URL Memento: Wayback Machine]

Irish Parliamentary Debates

Dáil Éireann. (1928). Ceisteanna—Questions. Oral Answers – Manuscripts Commission. – Dáil Éireann (6th Dáil) – Wednesday, 17 Oct 1928, Vol. 26 No. 4. Houses of the Oireachtas, Ireland. Retrieved 2022-10-24, from https://www.oireachtas.ie/en/debates/debate/dail/1928-10-17/3. [URL Memento: Wayback Machine]

Dáil Éireann. (2019). Departmental Websites – Dáil Éireann Debate – Tuesday 26 February 2019, Questions (7, 8, 9). Houses of the Oireachtas, Ireland. Retrieved 2019-10-24, from https://www.oireachtas.ie/en/debates/question/2019-02-26/9/. [URL Memento: Wayback Machine]

Dáil Éireann. (2021). Digital Archiving - Dáil Éireann Debate – Thursday 9 September 2021, Questions (267). Houses of the Oireachtas, Ireland. Retrieved 2022-09-14, from https://www.oireachtas.ie/en/debates/question/2021-09-09/267. [URL Memento: Wayback Machine]

Dáil Éireann. (2021). Intellectual Property – Dáil Éireann Debate – Thursday, 11 Nov 2021, Questions (175). Houses of the Oireachtas, Ireland. Retrieved 2022-09-29, from https://www.oireachtas.ie/en/debates/question/2021-11-11/175. [URL Memento]

Joint Committee on Tourism, Culture, Arts, Sport and Media. (2021a). Engagement with Chairperson Designate of the Board of the National Library of Ireland – Joint Committee on Tourism, Culture, Arts, Sport and Media debate – Wednesday, 13 Oct 2021 (Ireland). Houses of the Oireachtas, Ireland. Retrieved 2022-01-18, from https://www.oireachtas.ie/en/debates/debate/joint_committee_on_tourism_culture_arts_sport_and_media/2021-10-13/2/. [URL Memento: Archive.today]

Joint Committee on Tourism, Culture, Arts, Sport and Media. (2021b). Key Priorities and Legislation of the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media: Discussion – Joint Committee on Tourism, Culture, Arts, Sport and Media debate – Wednesday, 24 Nov 2021 (Ireland). Houses of the Oireachtas, Ireland. Retrieved 2022-04-26, from https://www.oireachtas.ie/en/debates/debate/joint_committee_on_tourism_culture_arts_sport_and_media/2021-11-24/3. [URL Memento: Wayback Machine]

Joint Committee on Tourism, Culture, Arts, Sport and Media. (2022). Future of Media Commission Report: Discussion Joint – Joint Committee on Tourism, Culture, Arts, Sport and Media debate – Wednesday, 14 Sep 2022. Retrieved 2022-09-29, from https://www.oireachtas.ie/en/debates/debate/joint_committee_on_tourism_culture_arts_sport_and_media/2022-09-14/2. [URL Memento: Wayback Machine]

Seanad Éireann. (2018). Copyright and Other Intellectual Property Law Provisions Bill 2018: Committee Stage – Wednesday, 3 Oct 2018, Vol. 260 No. 6. Houses of the Oireachtas. Retrieved 2021-12-24, from https://www.oireachtas.ie/en/debates/debate/seanad/2018-10-03/14 [URL Memento: Wayback Machine]

Seanad Éireann. (2021). An tOrd Gnó - Order of Business – Seanad Éireann Debate – Thursday, 30 Sep 2021, Vol. 278 No. 9. Houses of the Oireachtas, Ireland. Retrieved 2021-12-15, from https://www.oireachtas.ie/en/debates/debate/seanad/2021-09-30/8. [URL Memento: Wayback Machine]

Seanad Éireann. (2021). Nithe i dtosach suíonna - Commencement Matters – Digital Archiving - Seanad Éireann Debate – Tuesday, 23 Nov 2021, Vol. 280 No. 7. Houses of the Oireachtas, Ireland. Retrieved 2021-11-29, from https://www.oireachtas.ie/en/debates/debate/seanad/2021-11-23/3. [URL Memento: Wayback Machine]

Policy Documents

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. (2019). Maynooth University Research Ethics Policy. 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]

Screenshots

Beyond 2022. (n.d.). Virtual Tour: Record Treasury of Ireland. Beyond 2022. Retrieved 2022-10-31, from https://vrtour.virtualtreasury.ie/ [URL Memento: Archive.today]

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]

Department of Business, Enterprise and Innovation. (2017, September 5). Department of Business, Enterprise and Innovation – Home [Archived web page]. Department of Business, Enterprise and Innovation. Retrieved from https://web.archive.org/web/20170905050758/https://dbei.gov.ie/en/# [Web archive: Wayback Machine; source URL: https://dbei.gov.ie/en/#; Timestamp: 2017-09-05 05:07:58]

Department of Enterprise, Trade and Employment. (1999, October 4). Department of Enterprise, Trade and Employment – Home [Archived web page]. Department of Enterprise, Trade and Employment. Retrieved from https://web.archive.org/web/19991004144844/http://www.entemp.ie/ [Web archive: Wayback Machine; source URL: http://www.entemp.ie/; Timestamp: 1999-10-04 14:484:4]

Department of Enterprise, Trade and Employment. (2002, May 29). Department of Enterprise, Trade and Employment – Home [Archived web page]. Department of Enterprise, Trade and Employment. Retrieved from https://web.archive.org/web/20020529044054/http://www.entemp.ie:80/ [Web archive: Wayback Machine; source URL: http://www.entemp.ie:80/; Timestamp: 2002-05-29 04:40:54]

Department of Enterprise, Trade and Innovation. (2010, May 7). Department of Enterprise, Trade and Innovation – Home [Archived web page]. Department of Enterprise, Trade and Innovation. Retrieved from https://web.archive.org/web/20100507101638/http://www.deti.ie/ [Web archive: Wayback Machine; source URL: http://www.deti.ie/; Timestamp: 2010-05-07 10:16:38]

Department of Jobs, Enterprise and Innovation. (2011, June 6). Department of Jobs, Enterprise and Innovation – Home [Archived web page]. Department of Jobs, Enterprise and Innovation. Retrieved from https://web.archive.org/web/20110606063430/http://www.djei.ie/ [Web archive: Wayback Machine; source URL: http://www.djei.ie/; Timestamp: 2011-06-06 06:34:30]

Department of Jobs, Enterprise and Innovation. (2012, June 10). Department of Jobs, Enterprise and Innovation – Home [Archived web page]. Department of Jobs, Enterprise and Innovation. Retrieved from https://web.archive.org/web/20120610184500/http://enterprise.gov.ie/. [Web archive: Wayback Machine; source: http://enterprise.gov.ie/; Timestamp: 2012-06-10 18:45:00]

Department of Jobs, Enterprise and Innovation. (2012, June 13). Submissions Received 2012 on foot of the Copyright Review Consultation Paper [Archived web page]. Department of Jobs, Enterprise and Innovation. Retrieved from https://wayback.archive-it.org/org-

1444/20120613230622/http://www.djei.ie/science/ipr/crc_submissions2.htm [Web archive: Wayback Machine; source: URL: Timestamp: 2012-06-13 23:06:22]

Government of Ireland. (1996, December 24). Government of Ireland—Home [Archived web page]. Government of Ireland. Retrieved from https://web.archive.org/web/19961224221829/http://www.irlgov.ie/ [Web archive: Wayback Machine; source URL: http://www.irlgov.ie/;Timestamp: 1996-12-24 22:18:29]

Government of Ireland. (1997, January 5). Government of Ireland – Home [Archived web page]. Government of Ireland. Retrieved from https://web.archive.org/web/19970105013806/http://www.irlgov.ie/ [Web archive: Wayback Machine; source: URL: http://www.irlgov.ie/; Timestamp: 1997-01-05 01:38:06]

Government of Ireland. (2000, March 2). Government of Ireland – Home [Archived web page]. Government of Ireland - Home. Retrieved from https://web.archive.org/web/20000302062006/http://www.irlgov.ie/ [Web archive: Wayback Machine; source: URL: http://www.irlgov.ie/; Timestamp: 2000-03-02 06:20:06]

Government of Ireland. (2002, March 28). Government of Ireland – Home [Archived web page]. Government of Ireland. Retrieved from https://web.archive.org/web/20020328101202/http://www.irlgov.ie/ [Web archive: Wayback Machine; source: URL: http://www.irlgov.ie/; Timestamp: 2002-03-28 10:12:02]

Government of Ireland. (2008a, November 8). Government of Ireland – Home [Archived web page]. Government of Ireland. Retrieved from https://web.archive.org/web/20081108001743/http://www.irlgov.ie/ [Web archive: Wayback Machine; source: URL: http://www.irlgov.ie/; Timestamp: 2008-11-08 00:17:43]

Government of Ireland. (2008b, December 17). Government of Ireland – Home [Archived web page]. https://web.archive.org/web/20081217033949/http://www.gov.ie/en/ [Web archive: Wayback Machine; source: URL: http://www.gov.ie/en/; Timestamp: 2008-12-17 03:39:49]

Government of Ireland. (2011, July 3). Government of Ireland – Home. Retrieved from https://web.archive.org/web/20110703223040/http://www.gov.ie:80/en/ [Web archive: Wayback Machine; source: URL: http://www.gov.ie:80/en/; Timestamp: 2011-07-03 22:30:40]

International Internet Preservation Consortium. (2004, June 3). International Internet Preservation Consortium—Welcome (2004) [Archived web page]. International Internet Preservation Consortium. Retrieved from https://web.archive.org/web/20040603014115/http://netpreserve.org/about/index.php [Web archive: Wayback Machine; source URL: http://netpreserve.org/about/index.php; Timestamp: 2004-06-03 01:41:15]

National Library of Ireland. (n.d.). NLI Web Archive—Archive-It—Collections [Web page]. Archive-It. Retrieved 2022-07-04, from https://archive-it.org/home/nli/?show=Collections [URL Memento: Wayback Machine]

National Library of Ireland. (n.d.). NLI Web Archive—Archive-It—Sites [Web page]. Archive-It. Retrieved 2022-09-22, from https://archive-it.org/home/nli/?show=Sites [URL Memento: Wayback Machine]

Public Record Office of Northern Ireland (PRONI). (n.d). PRONI Web Archive [Web page]. NI Direct Government Services. Retrieved 2022-10-21, from https://webarchive.proni.gov.uk/#!/ [URL Memento: Wayback Machine]

UK Web Archive. (n.d.). UKWA Topics and Themes [Web page]. UK Web Archive. Retrieved 2022-06-19, from https://www.webarchive.org.uk/en/ukwa/category/ [URL Memento: Wayback Machine]

References

A | B | C | D | E | F

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]

Aasman, S. (2019). Finding Traces in YouTube's Living Archive: Exploring Informal Archival Practices. *TMG Journal for Media History*, 22(1), 35–55. DOI: 10.18146/tmg.435

Aasman, S., Haan, T. de, & Teszelszky, K. (2019). Web Archaeology: An Introduction. *TMG Journal for Media History*, 22(1), 1–5. DOI: 10.18146/tmg.433 [URL Memento: Wayback Machine]

Adam, A. (2007). *Implementing Electronic Document and Record Management Systems*. New York: Taylor & Francis.

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, [online], 21-22 September 2022*. 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: Digital Technology, Culture and Society*, 1(1–2), 146–159. DOI: 10.1080/24701475.2017.1309852

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]

Allport, G. W. (1954). *The Nature of Prejudice*. Cambridge, Massachusetts: Addison-Wesley Pub. Co. [Internet Archive]

American Library Association. (2008, February 21). Definitions of Digital Preservation [Web page]. Association for Library Collections & Technical Services (ALCTS). Retrieved 2019-02-06, from https://www.ala.org/alcts/resources/preserv/defdigpres0408. [URL Memento: Wayback Machine]

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

Antracoli, A., Duckworth, S., Silva, J., & Yarmey, K. (2014). Capture All the URLs: First Steps in Web Archiving. *Pennsylvania Libraries: Research & Practice*, 2(2), 155–170. DOI: 10.5195/palrap.2014.67 [URL Memento: Wayback Machine]

Archi, A. (2015). The Tablets of the Throne Room of the Royal Palace G of Ebla. *Archiv Für Orientforschung*, *53*, 9–18. [JSTOR ID: 44810781]

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]

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]

Ari, R. (2017). Importance and Role of Libraries in Our Society. *Tamralipta Mahavidyalaya Research Review: A Peer Reviewed National Journal of Interdisciplinary Studies*, *2*. Retrieved 2022-08-17, from https://tamraliptamahavidyalaya.org/tmrr/vol2/5RAri_F1.pdf . [URL Memento: Wayback Machine]

Arquivo.pt. (n.d.). Arquivo.pt Awards – sobre.arquivo.pt. Arquivo.Pt. Retrieved 2022-08-12, from https://sobre.arquivo.pt/en/collaborate/arquivo-pt-awards/. [URL Memento: Arquivo.pt]

Arvidson, A., Persson, K., & Mannerheim, J. (2000, August). The Kulturarw3 Project—The Royal Swedish Web Archiw3e—An example of 'complete' collection of web pages. *66th IFLA Council and General Conference Jerusalem, Israel, 13-18 August 2000*. IFLA Council and General Conference. Retrieved 2021-05-14, from

https://archive.ifla.org/IV/ifla66/papers/154-157e.htm. [URL Memento: Wayback Machine]

Ashforth, B. E., & Mael, F. (1989). Social Identity Theory and the Organization. *The Academy of Management Review*, *14*(1), 20–39. DOI: 10.2307/258189

Ashmore, R., Deaux, K., & McLaughlin-Volpe, T. (2004). An Organizing Framework for Collective Identity: Articulation and Significance of Multidimensionality. *Psychological Bulletin*, *130*(1), 80–114. DOI: 10.1037/0033-2909.130.1.80

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

Aturban, M. (2019a, 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]

Aturban, M. (2019b, September 25). Where did the archive go? Part 3: Public Record Office of Northern Ireland [Blog post]. Web Science and Digital Libraries Research Group. Retrieved 2021-06-22, from https://ws-dl.blogspot.com/2019/09/2019-09-25-where-did-archive-go-part-3.html. [URL Memento: Wayback Machine]

Aturban, M., Nelson, M. L., Weigle, M. C., Klein, M., & Van de Sompel, H. (2019). *Collecting 16K archived web pages from 17 public web archives* (arXiv:1905.03836). arXiv:1905.03836 [cs]

Aust, R. (2014, December 3). Online reactions to institutional crises: BBC Online and the aftermath of Jimmy Savile. In Publications Department [Paper]. Retrieved 2021-05-15, from https://sas-space.sas.ac.uk/6100/. [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]

Bansal, S., & Parmar, S. (2020). Decay of URLs Citation: A Case Study of Current Science. *Library Philosophy and Practice*, 01–09. Retrieved 2020-10-30, from https://digitalcommons.unl.edu/libphilprac/3582/. [URL Memento: Wayback Machine]

Barth, F. (Ed.). (1969). *Ethnic Groups and Boundaries: The Social Organization of Culture Difference*. Boston : Little, Brown and Co. Internet Archive, http://archive.org/details/ethnicgroupsboun0000unse

Bastian, J. (2001). A Question of Custody: The Colonial Archives of the United States Virgin Islands. *The American Archivist*, *64*(1), 96–114. DOI: 10.17723/aarc.64.1.h6k872252u2gr377

Bauer, R. (2018, August 2). What is the Difference Between Data Backup and Data Archive? [Blog post]. Backblaze Blog. Retrieved 2019-07-10, from https://www.backblaze.com/blog/data-backup-vs-archive/. [URL Memento: Wayback Machine]

Beal, V. (2010). The Difference Between The Internet And World Wide Web. Webopedia (Online/web). Webopedia. Retrieved 2022-02-11, from https://www.webopedia.com/insights/web-vs-internet/. [URL Memento: Wayback Machine]

Bearman, D. (1993). The Implications of 'Armstrong v. Executive of the President' for the Archival Management of Electronic Records. *The American Archivist*, 56(4), 674–689. JSTOR ID: 40293774

Beaudouin, V., Pehlivan, Z., & Stirling, P. (2019). Exploring the Memory of the First World War Using Web Archives: Web Graphs Seen from Different Angles. In N. Brügger & I. Milligan (Eds.), The SAGE Handbook of Web History (pp. 440–463). London: SAGE Publications.

Becker, C. (1938). What is Historiography? The American Historical Review, 44(1), 20–28. DOI: 10.2307/1840848

Behringer, W. (2006). Communications Revolutions: A Historiographical Concept. *German History*, *24*(3), 333–374. DOI: 10.1191/0266355406gh378oa

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

Bellamy, R. (2006). *Citizenship: A Very Short Introduction*. Oxford: Oxford University Press.

Ben-David, A. (2019). National web histories at the fringe of the web: Palestine, Kosovo, and the quest for online self-determination. In N. Brügger & D. Laursen (Eds.), *The Historical Web and Digital Humanities: The Case of National Web Domains* (pp. 89–109). London & New York: Routledge.

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.

Bergman, M. K. (2001). White Paper: The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing*, 7(1). DOI: 10.3998/3336451.0007.104 [URL Memento: Wayback Machine]

Berners-Lee, T. (1998). Hypertext Style: Cool URIs don't change. [Web page]. W3C. Retrieved 2019-03-08, from https://www.w3.org/Provider/Style/URI.html. [URL Memento: Wayback Machine]

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.).

Beyond 2022. (n.d.). Beyond 2022—Ireland's Virtual Record Treasury [Website]. Beyond 2022. Retrieved 2022-10-31, from https://beyond2022.ie/?page_id=2 [URL Memento: Archive.today]

Beyond 2022. (n.d.). Support [Web page] Beyond 2022. Retrieved 2022-10-31, from https://www.virtualtreasury.ie/support [URL Memento: Archive.today]

Beyond 2022. (n.d.). Virtual Tour: Record Treasury of Ireland. Beyond 2022. Retrieved 2022-10-31, from https://vrtour.virtualtreasury.ie/ [URL Memento: Archive.today]

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., Byrne, H., Lelkes-Rarugal, C., & Rossi, G. C. (2020, May 29). Using Webrecorder to archive UK political party leaders' social media after the UK General Election 2019 [Blog post]. UK Web Archive Blog. Retrieved 2021-06-19, from https://blogs.bl.uk/webarchive/2020/05/using-webrecorder-to-archive-uk-political-party-leaders-social-media-after-the-uk-general-election-2.html. [URL Memento: Wayback Machine]

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

Blizzard, S. M. (2006). *Women's Roles in the 1994 Rwanda Genocide and the Empowerment of Women in the Aftermath* [MA Thesis, Georgia Institute of Technology]. Retrieved from https://smartech.gatech.edu/handle/1853/11577. [URL Memento: Wayback Machine]

Bodenhausen, G., Kang, S. K., & Peery, D. (2012). Social categorization and the perception of social groups. In S. T. Fiske & C. N. Macrae (Eds.), *The SAGE Handbook of Social Cognition* (pp. 311–329). London: SAGE Publications Ltd.

Bødker, H., & Brügger, N. (2018). The shifting temporalities of online news: The Guardian's website from 1996 to 2015. *Journalism*, 19(1), 56–74. DOI: 10.1177/1464884916689153

Boston, G. (Ed.). (1998). Memory of the World: Safeguarding the documentary heritage, a guide to standards, recommended practices and reference literature related to the preservation of documents of all kinds. General Information Programme and UNISIST United Nations Educational, Scientific and Cultural Organization. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000112676. Document code: CII.98/WS/4

Bradsher, G. (2020, July 7). The Royal Archives of Ebla: Reference and Processing Archivists 4,000 Years Ago. *The Text Message*. Retrieved 2022-09-24, from https://text-

message.blogs.archives.gov/2020/07/07/the-royal-archives-of-ebla-reference-and-processing-archivists-4000-years-ago/

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

Brichford, M. (1989). The Provenance of Provenance in Germanic Areas. *Provenance, Journal of the Society of Georgia Archivists*, *7*(2). Retrieved from https://digitalcommons.kennesaw.edu/provenance/vol7/iss2/5

Brosius, M. (Ed.). (2003). *Ancient Archives and Archival Traditions: Concepts of Record-keeping in the Ancient World*. Oxford, New York: Oxford University Press.

Brown, A. (2006). Archiving Websites: A Practical Guide for Information Management Professionals. London: Facet Publishing.

Brubaker, R., & Cooper, F. (2000). Beyond Identity. *Theory and Society*, *29*, 1–47. DOI: 10.1023/A:1007068714468 [E-print: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/43651/?sequence=1]

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. (2012). When the Present Web is Later the Past: Web Historiography, Digital History, and Internet Studies. *Historical Social Research / Historische Sozialforschung*, 37(4 (142)), 102–117. JSTOR ID: 41756477

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. (2019). Understanding the Archived Web as Historical Source. In N. Brügger & I. Milligan (Eds.), *The SAGE Handbook of Web History* (pp. 16–29). London: SAGE Publications.

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. (2021a). Digital humanities and web archives: Possible new paths for combining datasets. *International Journal of Digital Humanities*, 2(1), 145–168. DOI: 10.1007/s42803-021-00038-z

Brügger, N. (2021b). Digital humanities and web archives: Possible new paths for combining datasets [Appendix]. OSF. Retrieved 2021-05-12, from https://doi.org/10.17605/OSF.IO/ZSU3D. [URL Memento: Wayback Machine]

Brügger, N. (2021c). 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.

Brügger, N. (2021d). The WARCnet network: The first year. *WARCnet Papers*. Aarhus, Denmark: WARCnet. Retrieved 2021-08-20, from https://cc.au.dk/fileadmin/user_upload/WARCnet/Bru__gger_The_first_year.pdf. [URL Memento: Wayback Machine]

Brügger, N. (2022). Tracing a historical development of conspiracy theory networks on the web: The hyperlink network of vaccine hesitancy on the Danish web 2006–2015. *Convergence*, 13548565221104988. DOI: 10.1177/13548565221104989

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., Laursen, D., & Nielsen, J. (2017). Exploring the domain names of the Danish web. In N. 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]

Brügger, N., Laursen, D., & Nielsen, J. (2019). Establishing a corpus of the archived web. In N. Brügger & D. Laursen (Eds.), *The Historical Web and Digital Humanities: The Case of National Web Domains* (pp. 124–142). 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]

Burner, M., & Kahle, B. (1996, September 15). Arc File Format [Web page]. Internet Archive. Retrieved 2021-12-18, from https://archive.org/web/researcher/ArcFileFormat.php. [URL Memento: Wayback Machine]

Bryan, D., & Gillespie, G. (2005). *Transforming Conflict: Flags and Emblems*. Institute of Irish Studies.

Byrne, H. (2019). Where are we now? A review of research on the history of women's soccer in Ireland. *Sport in History*, 39(2), 166–186. DOI: 10.1080/17460263.2019.1604422

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]

CAIN. (n.d.). PRONI Records on CAIN: About PRONI [Web page]. CAIN (Ulster University). https://cain.ulster.ac.uk/proni/aboutproni.html [URL Memento: Wayback Machine]

Carey, J. W. (1992). *Communication as Culture: Essays on Media and Society*. New York, London: Routledge. Internet Archive

Carter, R. G. S. (2006). Of Things Said and Unsaid: Power, Archival Silences, and Power in Silence. *Archivaria*, 215–233.

Castells, M. (2007). Communication, Power and Counter-power in the Network Society. *International Journal of Communication*, *1*(1), https://ijoc.org/index.php/ijoc/article/view/46

Castells, M. (2012). *Networks of outrage and hope: Social movements in the Internet age*. Chichester, UK: Wiley.

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: 'Opportunities, Challenges and Potentialities', (#EWAVirtual), Maynooth University Arts and Humanities Institute, Co. Kildare, Ireland, [online], 21-22 September 2022*. Retrieved 2020-09-22, from https://public.ccsds.org/outreach/DAIVideos.aspx. [URL Memento: Archive.today]

Cecco, L. (2023, July 6). Canadian judge rules thumbs-up emoji can represent contract agreement. *The Guardian*. Retrieved 2023-07-06, from https://www.theguardian.com/world/2023/jul/06/canada-judge-thumbs-up-emoji-sign-contract.

Chakrabarti, S., Dom, B., Kumar, S. R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Hypersearching the Web. *Scientific American*, 280(6), 54–60. http://www.jstor.org/stable/26058288. JSTOR

Charles, S. (2015, October 30). Private libraries of the rich and poor. *On History*. Retrieved from https://blog.history.ac.uk/2015/10/private-libraries-of-the-rich-and-poor/. [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]

Church, R. (1943). The Relationship Between Archival Agencies and Libraries. *The American Archivist*, *6*(3), 145–150. DOI: 10.17723/aarc.6.3.411852t15205l467

Cocciolo, A. (2015). The Rise and Fall of Text on the Web: A Quantitative Study of Web Archives. *Information Research: An International Electronic Journal*, 20(3). Retrieved 2022-01-19, from https://eric.ed.gov/?id=EJ1077827. [URL Memento: Wayback Machine]

Code, J. R., & Zaparyniuk, N. E. (2009). Social Identities, Group Formation, and the Analysis of Online Communities. In S. Hatzipanagos & S. Warburton (Eds.), *Handbook of Research on Social Software and Developing Community Ontologies* (pp. 86–101). Hershey, New York: Information Science Reference, IGI Global. DOI: 10.4018/978-1-60566-984-7.ch086.

Coenders, M., Gijsberts, M., Hagendoorn, L., & Scheepers, P. (2004). Introduction. In M. Gijsberts, L. Hagendoorn, A. Hagendoorn, & P. Scheepers (Eds.), *Nationalism and Exclusion of Migrants: Cross-national Comparisons* (pp. 1–25). Hanks, UK: Ashgate.

Cohen, A. P. (1985). *The Symbolic Construction of Community*. London, New York: Psychology Press.

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]

Collins, S. (2018). The National Library of Ireland. *Alexandria*, 28(3), 177–181. https://doi.org/10.1177/0955749019878523. DOI: 10.1177/0955749019878523

Committee on the Records of Government. (1985). Committee on the Records of Government. Report Undertaken by Council on Library Resources, Inc., Washington, DC.; Social Science Research Council, Washington, DC.; American Council of Learned Societies, New York. Retrieved 2019-01-30, from https://eric.ed.gov/?id=ED269018. [URL Memento: Wayback Machine] ERIC Number: ED269018

Computer Hope. (n.d.). TCP/IP [Web page]. Computer Hope. Retrieved 2021-12-09, from https://www.computerhope.com/jargon/t/tcpip.htm. [URL Memento: Wayback Machine]

Consortium of National and University Libraries. (n.d.). Consortium of National and University Libraries (CONUL) [Website]. CONUL. Retrieved 2022-07-07, from https://conul.ie/. [URL Memento: Wayback Machine]

Consortium of National and University Libraries. (2012). Submission by CONUL, In response to Copyright and Innovation: A Consultation Paper prepared by the Copyright Review

Committee, For the Department of Jobs, Enterprise and Innovation. Ireland: Consortium of National and University Libraries. Retrieved 2022-05-26, from https://enterprise.gov.ie/en/consultations/consultations-files/conul.pdf. [URL Memento: Wayback Machine]

Cooley, C. H. (1909). *Social organization; a study of the larger mind*. New York: Shocken Books. [Internet Archive]

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]

Coombs, L. A. (1989). A New Access System for the Vatican Archives. *The American Archivist*, *52*(4), 538–546. JSTOR

Copyright Review Committee. (2012). Copyright and Innovation: A Consultation Paper. Dublin, Ireland: Copyright Review Committee. Retrieved 2016-12-17, from http://www.cearta.ie/wp-content/uploads/2012/02/CRC-Consultation-Paper.pdf. [URL Memento: Wayback Machine]

Copyright Review Committee. (2013). Modernising Copyright [A Report prepared by the Copyright Review Committee for the Department of Jobs, Enterprise and Innovation]. Copyright Review Committee. Retrieved 2016-05-28, from http://www.cearta.ie/wp-content/uploads/2013/10/CRC-Report.pdf. [URL Memento: Wayback Machine]

Coram, R. G. (2015, July 24). Geo-location in the 2014 UK Domain Crawl. UK Web Archive Blog. Retrieved 2021-06-19, from https://blogs.bl.uk/webarchive/2015/07/geo-location-in-the-2014-uk-domain-crawl.html [URL Memento: Wayback Machine]

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, Denmark: 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]

Cowls, J. (2013, November 15). Fahrenheit 401: Digital Deletion Is Incompatible with Democracy. Josh Cowls. Retrieved 2020-12-27, from https://joshcowls.com/2013/11/15/fahrenheit-401-digital-deletion-is-incompatible-with-democracy/. [URL Memento: Wayback Machine]

Cox, R. J. (2001). *Managing Records as Evidence and Information*. Westwood, London: Greenwood Publishing Group.

Craigle, V., Retteen, A., & Keele, B. J. (2022). Ending Law Review Link Rot: A Plea for Adopting DOI. *Legal Reference Services Quarterly*, 0(0), 1–5. DOI: 10.1080/0270319X.2022.2089810

Crocetti, P. (2019, October 31). Archive vs. Backup and why you need to know the differences [Web page]. TechTarget. Retrieved 2022-01-29, from https://www.techtarget.com/searchdatabackup/tip/What-is-the-difference-between-archives-and-backups. [URL Memento: Wayback Machine]

Crowe, C. (2012, June 30). Ruin of Public Record Office marked loss of great archive. *The Irish Times*. Retrieved from https://www.irishtimes.com/opinion/ruin-of-public-record-office-marked-loss-of-great-archive-1.1069843 [URL Memento: Wayback Machine]

Cunningham, M. (1997a, March 17). Beware the ideas of March The Government Web site (http://www.irlgov.ie). *The Irish Times*, (COMPUTIMES), [CITY EDITION], p. 10. Proquest ID: 310308514

Cunningham, M. (1997b, January 27). Brewster's millions. *The Irish Times*, (COMPUTIMES), [CITY EDITION], p. 18. ProQuest ID: 310187147

Cunningham, M. (1997b, January 27). Brewster's millions. *The Irish Times*, (COMPUTIMES) [online] Retrieved from https://web.archive.org/web/19990117002422/http://www.irish-times.com/irish-times/paper/1997/0127/cmp1.html. [Wayback Machine, timestamp: 1999-01-17 00:24:22; URL source: http://www.irish-times.com/irish-times/paper/1997/0127/cmp1.html]

Cunningham, M. (1995, January 2). 1994: Year of the Net Michael Cunningham chronicles some of the main computing stories during annus mulllmedius. *The Irish Times*, (COMPUTIMES), [CITY EDITION], p. 17. ProQuest ID: 309949150

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]

Darcy, E., Charthaigh, C. N., Lalor, M., Sinnott, J., & Shank, C. (Eds.). (2021). In Her Shoes: Stories of the Eighth Amendment (Ireland). Digital Repository of Ireland, Collections. DOI: 10.7486/DRI.wm11nd02p

Dawson, C. (2019). A-Z of Digital Research Methods (Paperback). Oxon; New York: Routledge.

Day, M. (2003). *Collecting and preserving the World Wide Web: A feasibility study undertaken for the JISC and Wellcome Trust* (Version 1.0). UKOLN, University of Bath. https://web.archive.org/web/20030408060243/http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf. [Web archive: Wayback Machine; source URL: http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf; Timestamp: 2003-04-08 06:02:43]

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

Deaux, K. (2001). Social Identity. In *Encyclopedia of Women and Gender: Sex Similarities and Differences and the Impact of Society on Gender*, Volume 1 (pp. 1059–1067). Academic Press. Google-Books-ID: 7SXhBdqejgYC

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]

Delanty, G. (2009). *Community: 2nd edition* (2nd ed.). London: Routledge.

Dellavalle, R. P., Hester, E. J., Heilig, L. F., Drake, A. L., Kuntzman, J. W., Graber, M., & Schilling, L. M. (2003). Going, Going, Gone: Lost Internet References. *Science*, 302(5646), 787–788. American Association for the Advancement of Science. DOI: 10.1126/science.1088234

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

Denning, S. (2011, July 23). How Do You Change An Organizational Culture? Forbes. Retrieved 2014-11-16, from https://www.forbes.com/sites/stevedenning/2011/07/23/how-do-you-change-an-organizational-culture/. [URL Memento: Wayback Machine]

Department of Arts, Heritage, Regional, Rural and Gaeltacht Affairs. (2017a, April 21). Consultation on the Legal Deposit of published digital material in the 21st century in the context of Copyright legislation [Archived web page]. Department of Arts, Heritage, Regional, Rural and Gaeltacht Affairs, 21 April 2017. Retrieved from https://web.archive.org/web/20170424201229/http://www.ahrrga.gov.ie/consultation-on-the-legal-deposit-of-published-digital-material-in-the-21st-century-in-the-context-of-copyright-legislation/. [Web archive: Wayback Machine; source URL: http://www.ahrrga.gov.ie/consultation-on-the-legal-deposit-of-published-digital-material-in-the-21st-century-in-the-context-of-copyright-legislation/; Timestamp: 2017-04-24 20:12:29]

Department of Arts, Heritage, Regional, Rural and Gaeltacht Affairs. (2017b). Consultation on the Legal Deposit of published digital material in the 21st century in the context of Copyright legislation [Archived web page]. Department of Arts, Heritage, Regional, Rural and Gaeltacht Affairs, 21 April 2017. Retrieved from https://web.archive.org/web/20170424204921/http://www.ahrrga.gov.ie/app/uploads/2017/04/consultation-on-the-legal-deposit-of-published-digital-material-in-the-21st-century-in-the-context-of-copyright-legislation.pdf. [Web archive: Wayback Machine;

source URL: http://www.ahrrga.gov.ie/app/uploads/2017/04/consultation-on-the-legal-deposit-of-published-digital-material-in-the-21st-century-in-the-context-of-copyright-legislation.pdf; Timestamp: 2017-04-24 20:49:21]

Department of Culture, Heritage and the Gaeltacht. (2017, December). Department of Culture, Heritage, and the Gaeltacht response to submissions received in respect of the consultation on the Legal Deposit of published digital material in the 21st century in the context of Copyright Legislation. Department of Culture, Heritage and the Gaeltacht, December 2017. Retrieved from https://web.archive.org/web/20180305230900/https://www.chg.gov.ie/app/uploads/2017/12/public-consultation-reponse-summary-document-v2.pdf. [Web archive: Wayback Machine; source URL: https://www.chg.gov.ie/app/uploads/2017/12/public-consultation-reponse-summary-document-v2.pdf; Timestamp: 2018-03-05 23:09:00]

Department of Enterprise, Trade and Employment. (n.d.). Dara Calleary [Web page]. Department of Enterprise, Trade and Employment. Retrieved 2022-10-31, from https://enterprise.gov.ie/en/who-we-are/ministers/dara-calleary.html. [URL Memento: Wayback Machine]

Department of Enterprise, Trade and Employment. (n.d.). Robert Troy [Archived web page]. Department of Enterprise, Trade and Employment. Retrieved from https://web.archive.org/web/20220823133750/https://enterprise.gov.ie/en/Who-We-Are/Ministers/Robert-Troy/ [Web archive: Wayback Machine; source URL: https://enterprise.gov.ie/en/Who-We-Are/Ministers/Robert-Troy/; Timestamp: 2022-08-23 13:37:50

Department of Enterprise, Trade and Employment. (2016, August 4). General Scheme of a Copyright Bill approved by Government [Web page]. Department of Enterprise, Trade and Employment. Retrieved 2021-05-06, from https://enterprise.gov.ie/en/News-And-Events/Department-News/2016/August/04082016a.html. [URL Memento: Wayback Machine]

Department of Enterprise, Trade and Employment. (2021, December 5). Ministers—DETE [Archived web page]. Department of Enterprise, Trade and Employment. Retrieved from https://web.archive.org/web/20211205081704/https://enterprise.gov.ie/en/Who-We-Are/Ministers/. [Web archive: Wayback Machine; source URL: https://enterprise.gov.ie/en/Who-We-Are/Ministers/; Timestamp: 2021-12-05 08:17:04]

Department of Jobs, Enterprise and Innovation. (2012, June 13). Submissions Received 2012 on foot of the Copyright Review Consultation Paper [Archived web page]. Department of Jobs, Enterprise and Innovation. Retrieved from https://wayback.archive-it.org/org-1444/20120613230622/http://www.djei.ie/science/ipr/crc_submissions2.htm. [Web archive: NLI Web Archive; source: URL: Timestamp: 2012-06-13 23:06:22]

Department of Jobs, Enterprise, and Innovation. (2012, May 31). Consultation on the Review of the Copyright and Related Rights Act 2000 [Archived web page]. Department of Jobs,

Enterprise, and Innovation. Retrieved from https://wayback.archive-it.org/org-1444/20120613230629/http://www.djei.ie/science/ipr/copyright_review_2011.htm [Web archive: NLI Web Archive; source: URL: http://www.djei.ie/science/ipr/crc_submissions2.htm; Timestamp: 2012-06-13 23:06:22]

Department of Jobs, Enterprise, and Innovation. (2012, May 29). Copyright Review Committee announces limited extension of time for submissions [Archived web page]. Department of Jobs, Enterprise, and Innovation. Retrieved from https://web.archive.org/web/20120603221249/http://www.djei.ie/press/2012/20120529a.htm . [Web archive: Wayback Machine; source URL: http://www.djei.ie/press/2012/20120529a.htm; Timestamp: 2012-06-03 22:12:49]

Department of Jobs, Enterprise, and Innovation. (2012, March 2). Submissions Received by the Copyright Review Committee [Archived web page]. Department of Jobs, Enterprise, and Innovation. Retrieved from https://web.archive.org/web/20120618012622/http://www.djei.ie/science/ipr/crc_submissions.htm. [Web archive: NLI Web Archive; source: URL: http://www.djei.ie/science/ipr/crc_submissions2.htm; Timestamp: 2012-06-13 23:06:22]

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]

Dewey, J. (1916). *Democracy And Education*. New York: The MacMillan Company. [Internet Archive]

Digital Preservation Coalition. (n.d.). Web-archiving, Case study 1: The UK Web Archive [Web page]. Digital Preservation Coalition. Retrieved 2019-02-08, from https://www.dpconline.org/handbook/content-specific-preservation/web-archiving [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]

Digital Repository Ireland. (2022, September 16). DRI Project Archiving Reproductive Health Wins Digital Preservation Award. Digital Repository Ireland. https://www.dri.ie/dri-project-archiving-reproductive-health-wins-digital-preservation-award . [URL Memento: Wayback Machine]

Dollar, C. (1978). Appraising Machine-Readable Records. *The American Archivist*, 41(4), 423–430. https://doi.org/10.17723/aarc.41.4.g333h26662621363. DOI: 10.17723/aarc.41.4.g333h26662621363

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., & Heuvel, C. van den. (2009, April). Historical Infrastructures for Web Archiving: Annotation of Ephemeral Collections for Researchers and Cultural Heritage Institutions. *Massachusetts Institute of Technology 6th Media in Transition Conference, Cambridge, Massachusetts, 2009*. Massachusetts, USA. Retrieved from https://web.archive.org/web/20120616174354/http://web.mit.edu/comm-forum/mit6/papers/Dougherty_Heuvel.pdf [Web Archive: Wayback Machine; source URL: http://web.mit.edu/comm-forum/mit6/papers/Dougherty_Heuvel.pdf; Timestamp: 2012-06-16 17:43:54]

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]

Drucker, J. (2012). Humanistic Theory and Digital Scholarship. In Matthew K. Gold (Ed.), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press. DOI: 10.5749/minnesota/9780816677948.003.0011

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

Eldakar, Y., & Holownia, O. (2022, June 29). Republishing IIPC Collections Through Alternative Interfaces. *IIPC General Assembly and Web Archiving Conference, Library of Congress, May 2022* (IIPC WAC 2022). Retrieved from https://www.youtube.com/watch?v=15AGoOJBM6E [YouTube]

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]

Engholm, I. (2002). Digital style history: The development of graphic design on the Internet. *Digital Creativity*, 13(4), 193–211. DOI: 10.1076/digc.13.4.193.8672

Erskine, A. (2009). *A Companion to the Hellenistic World*. Maldon, Oxford, Victoria: John Wiley & Sons.

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]

Council Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the

information society, European Parliament, OJ L 167 (2001). https://eur-lex.europa.eu/eli/dir/2001/29/oj/eng [URL Memento: Archive.today] EUR-Lex Doc ID: 32001L0029

Council Directive 91/250/EEC of 14 May 1991 on the legal protection of computer programs, European Parliament, OJ 122 (1991). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A31991L0250&qid=1672866001086 [URL Memento: Wayback Machine] EUR-Lex Doc ID: 31991L0250

Council Directive 93/98/EEC of 29 October 1993 harmonizing the term of protection of copyright and certain related rights, European Parliament, OJ L 290 (1993). http://data.europa.eu/eli/dir/1993/98/oj/eng [URL Memento: Wayback Machine] EUR-Lex Doc ID: 31993L0098

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

Fanning, B. (2010). From Developmental Ireland to Migration Nation: Immigration and Shifting Rules of Belonging in the Republic of Ireland. *Economic and Social Review*, *41*(3), 395–412. Retrieved 2022-10-10, from http://www.ucd.ie/t4cms/soc_irishworkshop_economicandsocialreview10.pdf. [URL Memento: Wayback Machine]

Fanning, B., & Mutwarasibo, F. (2007). Nationals/non-nationals: Immigration, citizenship and politics in the Republic of Ireland. *Ethnic and Racial Studies*, *30*(3), 439–460. DOI: 10.1080/01419870701217506

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]

Feather, J. (1994). *Publishing, piracy, and politics: An historical study of copyright in Britain*. New York, N.Y: Mansell.

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

Finn, C. (2022a-03-14 19:14). McDonald says deletion of statements on SF website not attempt to pivot position on Russia. *The Journal*. https://www.thejournal.ie/mcdonald-russia-sinn-fein-ukraine-5710841-Mar2022/. [URL Memento: Wayback Machine]

Finn, C. (2022b-03-15 18:01). Taoiseach says removal of certain press statements from SF website is 'kind of Orwellian'. *The Journal*. https://www.thejournal.ie/sinn-fein-deleting-statements-taoiseach-5712160-Mar2022/. [URL Memento: Wayback Machine]

Fishbein, M. H. (1972). Appraising Information in Machine Language Form. *The American Archivist*, 35(1), 35–43. JSTOR ID: 40291594

Foot, K. A., & Schneider, S. M. (2006). *Web Campaigning*. Cambridge, MA, USA: MIT Press.

Fowler, S. (2017). Enforced Silences. In D. Thomas, S. Fowler, & V. Johnson (Eds.), *The Silence of the Archive* (pp. 1–39). London: Facet Publishing.

Gamson, W. A., Croteau, D., Hoynes, W., & Sasson, T. (1992). Media Images and the Social Construction of Reality. *Annual Review of Sociology*, *18*, 373–393. JSTOR

Gardner, J. B. (1997). Report on Documenting the Digital Age Conference. History Associates Incorporated. Retrieved from https://web.archive.org/web/20180524134523/https://www.historyassociates.com/wp-content/uploads/2015/09/Documenting-the-Digital-Age.pdf. [Web archive: Wayback Machine; source URL: Timestamp: 2018-05-24 13:45:23]

Gataveckaite, G. (2022, March 14). Sinn Féin deletes thousands of statements from its website due to 'outdated content'. *The Irish Independent*. https://www.independent.ie/irish-news/politics/sinn-fein-deletes-thousands-of-statements-from-its-website-due-to-outdated-content-41443385.html. [URL Memento: Wayback Machine]

Geoghegan, M. (2008). Social Movements, Community Development and the Discourse of National Community: An Analysis of Popular 'Progressive' Mobilisation in Ireland. In P. Herrmann (Ed.), *Governance and Social Professions: How Much Openness is Needed and how Much Openness is Possible?* (pp. 125–144). New York: Nova Science Publishers.

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]

Giddens, A. (1997). *Sociology*. Third edition. Cambridge: Polity Press.

Gillies, J., & Cailliau, R. (2000). *How the Web was Born: The Story of the World Wide Web*. Oxford, New York: Oxford University Press.

Gleason, P. (1983). Identifying Identity: A Semantic History. *The Journal of American History*, *69*(4), 910–931. DOI: 10.2307/1901196

Goh, D. H.-L., & Ng, P. K. (2007). Link decay in leading information science journals. *Journal of the American Society for Information Science and Technology*, 58(1), 15–24. DOI: 10.1002/asi.20513

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. (2021a). Preface. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The Past Web: Exploring Web Archives* (pp. xi–xiii). Cham, Switzerland: Springer.

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

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

Gooding, P., Terras, M., & Berube, L. (2019). Towards User-Centric Evaluation of UK Non-Print Legal Deposit: A Digital Library Futures White Paper. *Digital Library Futures*. https://digitalcommons.unl.edu/scholcom/180 [URL Memento: Wayback Machine]

Gooding, P., Terras, M., & Berube, L. (2021). Identifying the future direction of legal deposit in the United Kingdom: The Digital Library Futures approach. *Journal of Documentation*, 77(5), 1154–1172. DOI: 10.1108/JD-09-2020-0159

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

Graham, P. S. (1994). Intellectual Preservation: Electronic Preservation of the Third Kind. Commission on Preservation and Access. Retrieved 2021-05-07, from https://www.clir.org/pubs/reports/graham/intpres/. [URL Memento: Wayback Machine] ERIC Number: ED369414

Greene, D. (2020, September 21). Exploring Web Archive Networks: The Case of the 2018 Irish Presidential Election [Conference presentation (video; abstract)]. *Engaging with Web Archives: 'Opportunities, Challenges and Potentialities', (#EWAVirtual), Maynooth University Arts and Humanities Institute, Co. Kildare, Ireland, [online], 21-22 September 2022*. DOI: 10.5281/zenodo.4058013 [URL Memento: Wayback Machine] YouTube: https://www.youtube.com/watch?v=ld_fXefv-5E&t=1s

Greene, D., & Ryan, M. (2019, June 19). Exploring Selective Web Archives via Network Analysis: An Irish Case Study [Poster]. *LIBER Annual Conference, Dublin, Ireland, June 2019*. DOI: 10.5281/zenodo.3250157. [URL Memento: Wayback Machine]

Grotke, A. (2011, December). FEATURE: Web Archiving at the Library of Congress. Information Today. Retrieved 2018-07-21, from https://www.infotoday.com/cilmag/dec11/Grotke.shtml. [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]

Hayward, K. & Howard, K. (2002). 'Europeanisation And Hyphe-Nation: Renegotiating The Identity Boundaries Of Europe's Western Isles'. *Working Papers in British-Irish Studies*, No. 18. Retrieved 2022-01-22, from http://www.ucd.ie/ibis/filestore/18_khkh.pdf. [URL Memento: Wayback Machine]

Hayward, K. & Howard, K. (2007). Cherry-picking the Diaspora. In Fanning, B. (ed.) *Immigration and Social Change in the Republic of Ireland*. Manchester: Manchester University Press, pp 47–62.

Healy, S. (2016, October 27). Here today, gone tomorrow: A case study on the necessity for a more rigorous approach to the preservation of online Irish cultural and political heritage. Institutions and Ireland: Public Cultures, Trinity College Dublin, Ireland, 27 October 2016. Retrieved 2022-04-25, from https://hcommons.org/deposits/item/hc:45365/. DOI: 10.17613/xy5v-6b63. [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. (2021, April). Fleeting narratives: Web histories of Irish LGBT activism [Conference abstract]. *Women's History Association of Ireland Annual Conference 2020/2021:Besieged bodies: Gendered violence, sexualities and motherhood, University College Dublin, March 2021*. Retrieved 2021-02-26, from https://womenshistoryassociation.com/whai-conference-2020-2021/. [URL Memento: Wayback Machine]

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*. Aarhus, Denmark: WARCnet

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/]

Helmond, A. (2019). A Historiography of the Hyperlink: Periodizing the Web through the Changing Role of the Hyperlink. In N. Brügger & I. Milligan (Eds.), *The SAGE Handbook of Web History* (pp. 227–241). SAGE Publications.

Hendler, J. (2003). Science and the Semantic Web. *Science*, 299(5606), 520–521. DOI: 10.1126/science.1078874

Herman, E. S., & Chomsky, N. (1988). *Manufacturing Consent: The Political Economy of the Mass Media*. New York: Pantheon Books.

Hester, E. J., Heilig, L. F., Drake, A. L., Johnson, K. R., Vu, C. T., Schilling, L. M., & Dellavalle, R. P. (2004). Internet Citations in Oncology Journals: A Vanishing Resource? *Journal of the National Cancer Institute*, 96(12), 969–971. DOI: 10.1093/jnci/djh181

Hindley, M. (2013, August). The Rise of the Machines. *HUMANITIES: The Magazine of The National Endowment for the Humanities*, 34(4). Retrieved 2017-01-26, from https://www.neh.gov/humanities/2013/julyaugust/feature/the-rise-the-machines

Hockey, S. (2004). The History of Humanities Computing. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A Companion to Digital Humanities* (Web version). Oxford: Blackwell Publishing Professional. Retrieved 2018-06-28, from http://digitalhumanities.org:3030/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-2-1&toc.depth=1&toc.id=ss1-2-1&brand=9781405103213_brand . [URL Memento: Wayback Machine]

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: The Journal of National and International Library and Information Issues*, 25(11), 113–127. DOI: 10.7227/ALX.0023

Hofheinz, A. (2010). A History of Allah.com. In N. Brügger (Ed.), Web History (pp. 105–135). New York: Peter Lang.

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.

Holzmann, H., Nejdl, W., & Anand, A. (2016). The Dawn of Today's Popular Domains: A Study of the Archived German Web over 18 Years. *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pp. 73–82. DOI: 10.1145/2910896.2910901

Horne, E. (2016, November 5). The great flood of Florence, 50 years on. *The Guardian*. https://www.theguardian.com/artanddesign/2016/nov/05/the-great-flood-of-florence-50-years-on [URL Memento: Wayback Machine]

Howard, K. (2016). National Identity, Moral Panic and European Folk Devils. In B. Fanning & R. Munck (Eds.), *Globalization, Migration and Social Transformation: Ireland in Europe and the World* (pp. 169–182). London & New York: Routledge.

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

Hug, C. (2001). Moral Order and the Liberal Agenda in the Republic of Ireland. *New Hibernia Review / Iris Éireannach Nua*, *5*(4), 22–41. JSTOR

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

.IE Domain Registry. (2021). .IE Domain Profile Report 2021. .IE Domain Registry. Retrieved 2022-06-30, from https://www.weare.ie/wp-content/uploads/2022/01/IE-DPR-2021.pdf. [URL Memento: Wayback Machine]

Inkster, C. M. (1983). Geographically misplaced archives and manuscripts. *Archives & Manuscripts*, 113–124. Retrieved from, https://publications.archivists.org.au/index.php/asa/article/view/7559 [URL Memento: Wayback Machine]

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.). About the 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 [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 [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 [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.). IIPC WAC 2022 programme [Web page]. International Internet Preservation Consortium. Retrieved 2022-05-20, from https://netpreserve.org/ga2022/wac/. [URL Memento: Wayback Machine]

International Internet Preservation Consortium. (n.d.). IIPC webinar: Web Archiving the War in Ukraine. IIPC. Retrieved from https://netpreserve.org/mec-events/iipc-webinar-web-archiving-the-war-in-ukraine/

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

International Internet Preservation Consortium. (n.d.). International Internet Preservation Consortium - Archive-It [Web page]. Archive-It. Retrieved 2022-06-03, from https://archive-it.org/home/IIPC. [URL Memento: Wayback Machine]

International Internet Preservation Consortium. (n.d.). International Internet Preservation Consortium—Welcome (2004). International Internet Preservation Consortium. Retrieved from https://web.archive.org/web/20040603014115/http://netpreserve.org/about/index.php. [Web archive: Wayback Machine; source URL: http://netpreserve.org/about/index.php; Timestamp: 2004-06-03 01:41:15]

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.b). The WARC Format 1.0 [Web page]. IIPC/Github.Io. Retrieved 2021-07-06, from https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.0/. [URL Memento: Wayback Machine]

International Internet Preservation Consortium. (n.d.). Tools & software [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. (n.d.). Training materials [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 [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]

International Internet Preservation Consortium. (2017). Awesome Web Archiving (iipc/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]

International Internet Preservation Consortium - curators email list. (2017, April 4). Call for Participation: Archives Unleashed 4.0: Web Archive Datathon [distribution list communication].

Internet Archive Help Center. (n.d.). Wayback Machine General Information [Web page]. Internet Archive Help Center. Retrieved 2022-08-21, from https://help.archive.org/help/wayback-machine-general-information/. [URL Memento: Wayback Machine]

Irish Manuscripts Commission. (n.d.). *Mission* [Web page]. Irish Manuscripts Commission. Retrieved 2022-09-26, from https://www.irishmanuscripts.ie/about-us/mission/. [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. (2015a, April 27). Ten years of the UK web archive: What have we saved? [Conference presentation (slides)]. International Internet Preservation Consortium, General Assembly, Palo Alto, 27 April 2015, Palo Alto. Retrieved 2020-12-05, from https://blogs.bl.uk/webarchive/2015/09/ten-years-of-the-uk-web-archive-what-have-we-saved.html. [URL Memento: Wayback Machine]

Jackson, A. (2015b, 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]

James, J. (2019, June 7). The National Archives—Why Archives are for Everyone. The National Archives Blog. https://blog.nationalarchives.gov.uk/why-archives-are-for-everyone/. [URL Memento: Wayback Machine]

Jansma, R. (2020). Scoops and Brushes for Software Archaeology: Metadata Dating [MA Thesis: 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

Jenkins, R. (2008a). *Rethinking Ethnicity*. Second Edition. London: SAGE Publications Ltd

Jenkins, R. (2008b). *Social Identity*. London & New York: Routledge.

Jones, S. M., Klein, M., Sompel, H. V. de, Nelson, M. L., & Weigle, M. C. (2021). Interoperability for Accessing Versions of Web Resources with the Memento Protocol. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The Past Web: Exploring Web Archives* (pp. 101–126). Cham: Springer International Publishing.

Jurik, B., & Zierau, E. (2017). Data Management of Web Archive Research Data. *RESAW/IIPC Conference, School of Advanced Study, University of London, 14-16 June 2017*. RESAW/IIPC Conference, School of Advanced Study, University of London. Retrieved from https://web.archive.org/web/20210506225144/https://archivedweb.blogs.sas.ac.uk/files/2017/06/RESAW2017-JurikZierau-Data_management_of_web_archive_research_data.pdf [Web archive: Wayback Machine; source URL: https://archivedweb.blogs.sas.ac.uk/files/2017/06/RESAW2017-JurikZierau-Data_management_of_web_archive_research_data.pdf; Timestamp: 2021-05-0622:51:44]

Kahn, R. (2019). The Nation is in the network. In N. Brügger & D. Laursen (Eds.), *The Historical Web and Digital Humanities: The Case of National Web Domains* (pp. 161–177). London & New York: Routledge.

Kamalipour, Y. (Ed.). (1997). *The U.S. Media and the Middle East: Image and Perception*. United States: Greenwood Publishing Group.

Kelle, U. (1995). Introduction: An Overview of Computer-aided Methods in Qualitative Research. In U. Kelle, Geraldine Prein, & K. Bird (Eds.), *Computer-Aided Qualitative Data Analysis: Theory, Methods and Practice* (pp. 1–18). London: SAGE Publications.

Kellow, C., & Steeves, H. (1998). The role of radio in the Rwandan genocide. *Journal of Communication*, *48*(3), 107–128. DOI: 10.1111/j.1460-2466.1998.tb02762.x

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.DOI: 10.1016/S1464-9055(00)00178-0

Klein, M., Sompel, H. V. de, Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., & Tobin, R. (2014). Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLoS ONE*, 9(12), e115253. DOI: 10.1371/journal.pone.0115253

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]

Koehler, W. (2015). *Ethics and Values in Librarianship: A History*. New York & London: Rowman & Littlefield.

Koerbin, P. (2013a, July 2). Web Archiving in a Fast Moving World: The Tricky Task of Capturing Websites While the News Is Hot [Blog post]. National Library of Australia Blog. Retrieved from https://web.archive.org/web/20150910044329/https://www.nla.gov.au/australias-web-archives/2013/07/02/web-archiving-in-a-fast-moving-world. [Web archive: Wayback Machine; source: URL; Timestamp: 2015-09-10 04:43:29

Koerbin, P. (2013b, August 9). Archiving Online Election Campaigns: Explaining the Process Behind How We Preserve Australia's Election Websites [Blog post]. National Library of Australia Blog. Retrieved from https://web.archive.org/web/20150905084451/https://www.nla.gov.au/australias-web-archives/2013/08/09/archiving-online-election-campaigns [Web archive: Wayback Machine; source URL: https://www.nla.gov.au/australias-web-archives/2013/08/09/archiving-online-election-campaigns; Timestamp: 2015-09-05 08:44:51

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.

Kornprobst, M. (2005). Episteme, nation-builders and national identity: The re-construction of Irishness. *Nations & Nationalism*, *11*(3), 403–421. DOI: 10.1111/j.1354-5078.2005.00211.x

Kuny, T. (1997, September 4). A Digital Dark Ages? Challenges in the Preservation of Electronic Information. *63rd IFLA General Conference, Copenhagen, August 31 - September 5, 1997*. Retrieved 2021-07-23, from http://archive.ifla.org/IV/ifla63/63kuny1.pdf. [URL Memento: Wayback Machine]

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

Lancaster, F. W. (1995). The Evolution of Electronic Publishing. *Library Trends*, 43(4). Retrieved 2022-07-26, from https://www.ideals.illinois.edu/items/7940 . [URL Memento: Wayback Machine] Permalink: https://hdl.handle.net/2142/7981

Larivière, J. (2000). *Guidelines for legal deposit legislation* (UNESCO) [Programme and meeting document]. United Nations Educational, Scientific and Cultural Organization. Retrieved 2019-04-25, from https://unesdoc.unesco.org/ark:/48223/pf0000121413. Document code: CII.00/WS/7

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]

Lawrence, S., Pennock, D. M., Flake, G. W., Krovetz, R., Coetzee, F. M., Glover, E., Nielsen, F. Å., Kruger, K., & Giles, C. L. (2001). Persistence of Web References in Scientific Research. *Computer*, 34(1), 26–31. IEEE Xplore. DOI: 10.1109/2.901164

Lay, P. (2017, May 5). History has a History. *History Today*, 67(5). Retrieved 2020-09-04, from https://www.historytoday.com/archive/editor/history-has-history [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]

Leiner, B. M., Cerf, V. G., Clark, D. D., Kahn, R. E., Kleinrock, L., Lynch, D. C., Postel, J., Roberts, L. G., & Wolff, S. (1997). *Brief History of the Internet*. Internet Society. Retrieved 2018-03-29, from https://cdn.prod.internetsociety.org/wp-content/uploads/2017/09/ISOC-History-of-the-Internet_1997.pdf. [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]

Leetaru, K. (2019, May 7). Why Web Archives Need To Engage With Researchers. *Forbes*, AI & Big Data . Retrieved 2019-11-30, from https://www.forbes.com/sites/kalevleetaru/2019/05/07/why-web-archives-need-to-engage-with-researchers/. [URL Memento: Wayback Machine]

Lentin, R. (1998). 'Irishness', the 1937 Constitution, and Citizenship: A Gender and Ethnicity View. *Irish Journal of Sociology*, *8*(1), 5–24. DOI: 10.1177/079160359800800101

Lialina, Olia, & 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]

Library of Congress. (n.d.). ARC_IA, Internet Archive ARC file format [Web page]. Sustainability of Digital Formats: Planning for Library of Congress Collections. Retrieved 2021-12-18, from https://www.loc.gov/preservation/digital/formats/fdd/fdd000235.shtml

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]

Library of Trinity College Dublin. (n.d.a). Electronic Legal Deposit: Accessing UK eLD content [Web page]. Trinity College Dublin. Retrieved 2022-09-28, from https://libguides.tcd.ie/c.php?g=691904&p=4957591. [URL Memento: Archive.Today]

Library of Trinity College Dublin. (n.d.b). Electronic Legal Deposit: History and Background [Web page]. Trinity College Dublin. Retrieved 2022-09-28, from https://libguides.tcd.ie/c.php?g=691904&p=4957589. [URL Memento: Wayback Machine]

Liulevicius, V. (2020, June 21). The Social Impact of the Printing Press. Wondrium Daily. https://www.wondriumdaily.com/the-social-impact-of-the-printing-press/. [URL Memento: Wayback Machine]

Lomborg, S. (2019). Ethical Considerations for Web Archives and Web History Research. In N. Brügger & I. Milligan (Eds.), *The SAGE Handbook of Web History* (pp. 99-111). London: SAGE Publications.

Lowcock, J. (2020, February 26). Guide: How to Delete Your Site from the Internet Archive (Wayback Machine / Archive.org) [Digital-media; personal]. Joshua Lowcock, New York. Retrieved 2020-05-26, from https://www.joshualowcock.com/tips-tricks/how-to-delete-your-site-from-the-internet-archive-wayback-machine-archive-org/. [URL Memento: Wayback Machine]

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]

Mackinnon, K. (2020). DELETE MY ACCOUNT: Ethical Approaches to Researching Youth Cultures in Historical Web Archives. *Engaging with Web Archives: 'Opportunities, Challenges and Potentialities', (#EWAVirtual), Maynooth University Arts and Humanities Institute, Co. Kildare, Ireland, [online], 21-22 September 2022*. Retrieved from https://www.youtube.com/watch?v=2lBX1Om6GLM

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: Digital Technology, Culture and Society*, 0(0), 1–16. DOI: 10.1080/24701475.2022.2051331

Macrotrends. (2017). Ireland Immigration Statistics 1960-2023. Macrotrends. Retrieved 2023-06-28, from https://www.macrotrends.net/countries/IRL/ireland/immigration-statistics. [URL Memento: Wayback Machine]

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]

Maguire, M., & Cassidy, T. M. (2009). The New Irish Question: Citizenship, Motherhood and the Politics of Life Itself. *Irish Journal of Anthropology*, *12*(3), Retrieved from https://mural.maynoothuniversity.ie/2851/. [URL Memento: Wayback Machine]

Maguire, M. (2022). 16 January 1922: The 'Surrender of Dublin Castle' [Web page]. RTÉ: Century Ireland. Retrieved 2022-05-15, from https://www.rte.ie/centuryireland/index.php/articles/16-january-1922-the-surrender-of-dublin-castle [URL Memento: Wayback Machine]

Malone, D. (n.d.). Early Irish Web Stuff [Web page]. TCD Maths. Retrieved 2017-05-16, from https://www.maths.tcd.ie/~dwmalone/early-web.html. [URL Memento: Wayback Machine]

Markwell, J., & Brooks, D. W. (2002). Broken Links: The Ephemeral Nature of Educational WWW Hyperlinks. *Journal of Science Education and Technology*, 11(2), 105–108. DOI: 10.1023/A:1014627511641

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.

Matthews, B. (1890) The Evolution of Copyright. *Political Science Quarterly*, 5(4), pp 583-602. DOI: https://www.jstor.org/stable/2139530.

Maurer, Y. (2022, August 10). Investigate holdings of web archives through summaries: Cdx-summarize. Retrieved 2022-08-10, from https://netpreserveblog.wordpress.com/2022/08/10/investigate-holdings-of-web-archives-through-summaries-cdx-summarize/. [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]

McChesney, R., & Schiller, D. (2003). The Political Economy of International Communications Foundations for the Emerging Global Debate about Media Ownership and Regulation (Technology, Business and Society Programme Paper Number 11). United Nations Research Institute for Social Development. Retrieved 2022-04-03, from https://www.files.ethz.ch/isn/90745/11.pdf. [URL Memento: Wayback Machine]

McDonald, H. (2008, April 2). Irish prime minister Ahern resigns amid financial controversy. *The Guardian*. Retrieved 2021-10-23, from https://www.theguardian.com/world/2008/apr/02/ireland. [URL Memento: Wayback Machine]

McGee, H. (2018, December 28). Move to begin in 2019 to release State papers after 20 years. *The Irish Times*. Retrieved 2018-12-28, from https://www.irishtimes.com/news/politics/move-to-begin-in-2019-to-release-state-papers-after-20-years-1.3742288 [URL Memento: Wayback Machine]

McKemmish, S. (2005). Traces: Document, archives record, archive. In S. McKemmish, M. Piggott, B. Reed, & F. Upward (Eds.), *Archives: Recordkeeping in Society* (pp. 1–20). Wagga Wagga, N.S.W.: Centre for Information Studies, Charles Sturt University.

McKernan, L. (n.d.). A Short History of Film Archiving. Retrieved 2021-05-06, from http://www.bftv.ac.uk/events/archhist.htm. [URL Memento: Wayback Machine]

McLuhan, M. (Director). (1960). The Communications Revolution. Ohio State University. http://www.marshallmcluhanspeaks.com/panel/1960-the-communications-revolution/. [URL Memento: Wayback Machine]

McLuhan, M. (1967). *The Medium Is the Massage: An Inventory of Effects*. New York, London, Toronto: Bantam Books. [Internet Archive]

McLuhan, M. (1970). Living in an Acoustic World. Marshall McLuhan Speaks Special Collection, University of South Florida Public Lecture Retrieved from http://www.marshallmcluhanspeaks.com/lecture/1970-living-in-an-acoustic-world/. [URL Memento: Wayback Machine]

Meehan, H. (1996). Beneath the skin and between the ears: A case study in the politics of representation. In S. Chaiklin & J. Lave (Eds.), *Understanding Practice: Perspectives on Activity and Context* (pp. 241–268). Cambridge: Cambridge University Press.

Media Texthack Team. (2014). *Media Studies 101*. BCcampus. [B.C. Open Textbook Collection]

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

Meyer, E. T., Yasseri, T., Halt, S. A., Josh Cowls, Schroeder, R., & Helen Margetts. (2017). Analysing the UK web domain and exploring 15 years of UK universities on the web. In N. Brügger & R. Schroeder (Eds.), The Web as History: Using Web Archives to Understand the Past and the Present (Online/pdf, pp. 23–44). London: UCL Press. DOI: 10.14324/111.9781911307563 [URL Memento: Wayback Machine]

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]

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

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]

Millward, G. (2015). Digital barriers and the accessible web: Disabled people, information and the internet (Big UK Domain Data for the Arts and Humanities) [Case study]. School of Advanced Study, University of London. Retrieved 2022-01-25, from http://sas-space.sas.ac.uk/6104/ [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]

Mirtaheri, S. M., Dinçtürk, M. E., Hooshmand, S., Bochmann, G. V., Jourdan, G.-V., & Onut, I. V. (2013). A brief history of web crawlers. *Proceedings of the 2013 Conference of the Center for Advanced Studies on Collaborative Research*, 40–54. DOI: 10.5555/2555523.2555529; Preprint: http://arxiv.org/abs/1405.0749

Mohr, G., Michael F. Stack, Igor Ranitovic, Daniel Avery, & Michele Kimpton. (2004). Introduction to Heritrix, an archival quality web crawler. *Proceedings of the 4th International Web Archiving Workshop (IWAW'04)*, Bath, UK, July 2004, pp. 1–15. Retrieved 2014-07-18, from https://www.bibsonomy.org/bibtex/209d70d4ea1810fe89522755a0982169f/jaeschke. [URL Memento: Wayback Machine]

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]

Morris, J. W. (2019). Hearing the Past: The Sonic Web from MIDI to Music Streaming. In N. Brügger & I. Milligan (Eds.), *The SAGE Handbook of Web History* (pp. 491–504). London: SAGE Publications.

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]

Muir, A. (2005). Legal deposit of digital publications [PhD Dissertation, Loughborough University]. Retrieved 2022-02-25, from https://repository.lboro.ac.uk/articles/thesis/Legal_deposit_of_digital_publications/9415538/1. uk.bl.ethos.444005

Mukherji, P. N. (2010). Civic-secular and ethnic nationalisms as bases of the nation-state: Multiculturalism at the crossroads? *Asian Ethnicity*, *11*(1), 1–23. DOI: 10.1080/14631360903506745

Muldoon, J. (1997). Communications Revolutions: Little History, Much Myth. In J. B. Gardner (Ed.), Report on Documenting the Digital Age Conference. Retrieved from https://web.archive.org/web/20180524134523/https://www.historyassociates.com/wp-content/uploads/2015/09/Documenting-the-Digital-Age.pdf. [Web archive: Wayback Machine; source URL: Timestamp: 2018-05-24 13:45:23]

Munck, R. (2011). Ireland in the world, the world in Ireland. In B. Fanning (Ed.), *Globalisation, Migration and Social Transformation: Ireland in Europe and the World*. Ashgate Publishing Limited.

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

Murchan, R. (2020a-09). PRONI Web Archive: A Collaborative Approach [Conference presentation (video; abstract)]. Engaging with Web Archives: 'Opportunities, Challenges and Potentialities', (#EWAVirtual), 21-22 September 2020, Maynooth University Arts and Humanities Institute, Co. Kildare, Ireland, [online]. https://www.youtube.com/watch?v=OT_DXyhUCD4&feature=youtu.be [YouTube] DOI: 10.5281/zenodo.4058013

Murchan, R. (2020b-10-28). PRONI Web Archive: A Collaborative Approach (Blog) [Blog post]. UK Web Archive Blog. Retrieved 2022-01-08, from https://blogs.bl.uk/webarchive/2020/10/proni-web-archive-a-collaborative-approach.html [URL Memento: Wayback Machine]

Musso, M., & Merletti, F. (2016). This is the future: A reconstruction of the UK business web space (1996–2001). New Media & Society, 18(7), 1120–1142. DOI: 10.1177/1461444816643791

N | O | P | Q | R | S

Nanni, F. (2017). Reconstructing a website's lost past Methodological issues concerning the history of Unibo.it. *Digital Humanities Quarterly*, 011(2). Retrieved 2018-08-06, from http://www.digitalhumanities.org/dhq/vol/11/2/000292/000292.html. [URL Memento: Wayback Machine]

National Archives of Ireland. (n.d.). News: Shaping Our Future in the Information Age – National Archives Strategic Plan 2021-2025 [Web page]. National Archives of Ireland (NAI). Retrieved 2022-05-20, from https://www.nationalarchives.ie/news/shaping-our-future-in-the-information-age-national-archives-strategic-plan-2021-2025/. [URL Memento: Wayback Machine]

National Archives of Ireland. (n.d.). Public Record Office of Ireland: The Story of a Building— 2022 Commemoration Programme [Web page]. National Archives of Ireland (NAI). Retrieved 2022-07-05, from https://www.nationalarchives.ie/2021commemorationprogramme/public-record-office-of-ireland-the-story-of-a-building-exhibition/. [URL Memento: Wayback Machine]

National Archives of Ireland. *Report of the Director of the National Archives for 2012*. National Archives of Ireland. Retrieved 2019-12-10, from https://www.nationalarchives.ie/wp-content/uploads/2019/03/DirectorsReport2012.pdf. [URL Memento: Wayback Machine]

National Archives of Ireland. *Report of the Director of the National Archives for 2013. National Archives of Ireland*. Retrieved 2019-12-10, from https://www.nationalarchives.ie/wp-content/uploads/2019/03/DirectorsReport2014.pdf. [URL Memento: Wayback Machine]

National Archives of Ireland. *Report of the Director of the National Archives for 2014*. National Archives of Ireland. Retrieved 2019-12-10, from https://www.nationalarchives.ie/wp-content/uploads/2019/03/DirectorsReport2014.pdf. [URL Memento: Wayback Machine]

National Archives of Ireland. *Report of the Director of the National Archives for 2015*. National Archives of Ireland. Retrieved 2019-12-10, from https://www.nationalarchives.ie/wp-content/uploads/2019/03/DirectorsReport2015.pdf. [URL Memento: Wayback Machine]

National Archives of Ireland. *Report of the Director of the National Archives for 2016*. National Archives of Ireland. Retrieved 2019-12-10, from https://www.nationalarchives.ie/wp-content/uploads/2019/03/National-Interactive-Archives-Report.pdf. [URL Memento: Wayback Machine]

National Archives of Ireland. *Report of the Director of the National Archives for 2017*. National Archives of Ireland. Retrieved 2019-12-10, from https://www.nationalarchives.ie/wp-content/uploads/2019/06/18-092-National-Archives-Report-08.01.19_V08-002.pdf. [URL Memento: Wayback Machine]

National Archives of Ireland. *Report of the Director of the National Archives for 2018*. National Archives of Ireland. Retrieved 2019-12-10, from https://www.nationalarchives.ie/wp-content/uploads/2021/05/NA_2018-Annual-Report-redux.pdf. [URL Memento: Wayback Machine]

National Archives of Ireland. *Report of the Director of the National Archives for 2019*. National Archives of Ireland (NAI). Retrieved 2021-09-15, from https://www.nationalarchives.ie/wp-content/uploads/2021/07/21.03.15-NA-AR-2019-Redux.pdf. [URL Memento: Wayback Machine]

National Archives of Ireland. *Report of the Director of the National Archives for 2020*. National Archives of Ireland. Retrieved 2022-03-02, from https://www.nationalarchives.ie/wp-content/uploads/2022/01/21.10.01-NA-AR-2020-English-and-Gaeilge-V001_Directors-Exp-compressed_2.pdf. [URL Memento: Wayback Machine]

National Archives of Ireland. (2021-03-11). *Shaping our future in the Information Age: The National Archives 2021—2025 Strategic Plan*. National Archives of Ireland. Retrieved 2022-03-02, from https://www.nationalarchives.ie/wp-content/uploads/2021/03/NA-Strategic-Plan-2021-25_FINAL_RSZ.pdf. [URL Memento: Wayback Machine]

National Digital Stewardship Alliance. (2022, April 14). About the NDSA [Web page]. National 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 Ireland. (2009). *Collection Development Policy 2009-2011*. [Web archive]. Dublin, Ireland: National Library of Ireland (NLI). Retrieved from https://web.archive.org/web/20211125013643/https://www.nli.ie/GetAttachment.aspx?id=be415641-55bb-467c-aa45-50da5c389f78. [Web Archive: Wayback Machine; source URL: https://www.nli.ie/GetAttachment.aspx?id=be415641-55bb-467c-aa45-50da5c389f78; Timestamp: 2021-11-25 01:36:43]

National Library of Ireland. (2022). *Collection Development Policy 2022-2026*. Dublin, Ireland: National Library of Ireland (NLI). Retrieved from https://www.nli.ie/sites/default/files/2022-/nli_collections_revisedpolicy_2022.pdf. [URL Memento: Wayback Machine]

National Library of Ireland. (2011). *NLI Annual Report, 2011*. [Web archive]. Dublin, Ireland: National Library of Ireland (NLI). Retrieved from https://web.archive.org/web/20211125013643/https://www.nli.ie/GetAttachment.aspx?id=be415641-55bb-467c-aa45-50da5c389f78. [Web Archive: Wayback Machine; source URL: https://www.nli.ie/GetAttachment.aspx?id=be415641-55bb-467c-aa45-50da5c389f78; Timestamp: 2021-11-25 01:36:43]

National Library of Ireland. (2012). *NLI Annual Report, 2012*. [Web archive]. Dublin, Ireland: National Library of Ireland (NLI). Retrieved from https://web.archive.org/web/20211125013640/https://www.nli.ie/GetAttachment.aspx?id=f095220a-4e33-47cf-b64f-28850e05955d. [Web Archive: Wayback Machine; source URL: https://www.nli.ie/GetAttachment.aspx?id=f095220a-4e33-47cf-b64f-28850e05955d; Timestamp: 2021-11-25 01:36:40]

National Library of Ireland. (2014). *Legal Deposit 2014*. [Web Archive]. Dublin, Ireland: National Library of Ireland (NLI). Retrieved from https://web.archive.org/web/20211125015832/https://www.nli.ie/GetAttachment.aspx?id=e2a3e901-8860-4473-9d2b-77741b45f78f. [Web archive: Wayback Machine; source URL: https://www.nli.ie/GetAttachment.aspx?id=e2a3e901-8860-4473-9d2b-77741; Timestamp: 2021-11-25 01:58:32]

National Library of Ireland. (2022). About the Library. [Archived web page]. National Library of Ireland (NLI). Retrieved from https://web.archive.org/web/20220619021147/https://www.nli.ie/en/intro/about-the-library.aspx. [Web archive: Wayback Machine; source URL: https://www.nli.ie/en/intro/about-the-library.aspx; Timestamp: 2022-06-19 02:11:47]

National Library of Ireland. (2022). History of the Library. [Archived web page]. National Library of Ireland (NLI). Retrieved from https://web.archive.org/web/20221116002213/https://www.nli.ie/en/history-of-the-library.aspx [Web Archive: Wayback Machine; source URL: https://www.nli.ie/en/history-of-the-library.aspx; Timestamp: 2022-11-16 00:22:13]

National Library of Ireland. (2022). History of the Library: 1877 to 1926. [Archived web page]. National Library of Ireland (NLI). https://web.archive.org/web/20221116002213/https://www.nli.ie/en/history-of-the-

library-1877-1926.aspx. [Web archive: Wayback Machine: source URL: https://www.nli.ie/en/history-of-the-library-1877-1926.aspx; Timestamp: 2022-11-16 00:22:13]

National Library of Ireland. (2022). History of the Library: 1927 to Present. [Archived web page]. National Library of Ireland (NLI). https://web.archive.org/web/20220817171142/https://www.nli.ie/en/history-of-the-library-1927-to-present.aspx. [Web archive: Wayback Machine; source URL: https://www.nli.ie/en/history-of-the-library-1927-to-present.aspx; Timestamp: 2022-08-17 17:11:42]

National Library of Ireland. (2022). History of the Library: Origins. [Archived web page]. National Library of Ireland (NLI). https://web.archive.org/web/20220817170613/https://www.nli.ie/en/history-of-the-library-origins.aspx. [Web archive: Wayback Machine]

National Library of Ireland. (2022). Irish Domain Web Archive. National Library of Ireland (NLI). Retrieved 2022-09-26, from https://www.nli.ie/en/irish-domain-web-archive.aspx [URL Memento: Wayback Machine]

National Library of Ireland. (2022). Selective Web Archive Collections. National Library of Ireland (NLI). Retrieved 2022-09-29, from https://www.nli.ie/en/udlist/web-archive-collections.aspx [URL Memento: Wayback Machine]

National Library of Ireland. (2022). *NLI 2022—2026 Strategy*. Dublin, Ireland: National Library of Ireland (NLI). Retrieved from https://www.nli.ie/sites/default/files/2022-11/nlistrategyenglishweb2022.pdf. [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]

Nelson, M. L. (2018a, November 28). '44 days' is often quoted, and probably derives from @brewster_kahle's 1997 Scientific American article: 'Preserving the Internet' https://t.co/Nx8IoI8emy though note that his 1996 preprint of that article uses '75 days' https://t.co/zNnYOCrHXp #wdpd18 [Tweet]. Twitter @phonedude_mln. https://twitter.com/phonedude_mln/status/1067858148952891393. [Twitter]

Nelson, M. L. (2018b, November 28). '100 days' probably comes from this quote from Kahle in this 2003 WaPo article: 'On the Web, Research Work Proves Ephemeral' https://t.co/JpQbMX7Lei https://t.co/84GB6Nnx0z in both 1997 & 2003, note that the granularity is for URL or an individual page, not for 'site' #wdpd18 [Tweet]. @phonedude_mln. https://twitter.com/phonedude_mln/status/1067858149842055169 [Twitter]

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

Nic Craith, M. (2002). *Plural Identities—Singular Narratives: The Case of Northern Ireland*. New York and Oxford: Berghahn Books.

NIdirect. (2015, December 9). About the PRONI Web Archive. Nidirect Government Services. Retrieved 2021-12-17, from https://www.nidirect.gov.uk/articles/about-proni-web-archive [URL Memento: Wayback Machine]

NIdirect. (2016, March 4). Public Record Office of Northern Ireland (PRONI). Nidirect Government Services. Retrieved 2022-09-01, from https://www.nidirect.gov.uk/campaigns/public-record-office-northern-ireland-proni [URL Memento: Wayback Machine]

NIdirect. (n.d.). Getting to PRONI and opening hours. Nidirect Government Services. Retrieved 2021-06-16, from https://www.nidirect.gov.uk/articles/getting-proni-and-opening-hours [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]

Ntoulas, A., Cho, J., & Olston, C. (2004). What's New on the Web?: The Evolution of the Web from a Search Engine Perspective. Proceedings of the 13th International Conference on World Wide Web (WWW '04), 1–12. ACM Digital Library. DOI: 10.1145/988672.988674

O'Connell, H. (2022, March 13). Sinn Féin wipes years of media statements from website. *Sunday Independent*. Retrieved 2022-03-13, from https://www.independent.ie/irish-news/news/sinn-fein-wipes-years-of-media-statements-from-website-41440873.html. [URL Memento: Wayback Machine]

O'Dell, E. (2012, February 29). Copyright and Innovation – The CRC Consultation Paper [Blog post]. CEARTA.IE, the Irish for Rights. Retrieved 2020-11-28, from http://www.cearta.ie/2012/02/copyright-and-innovation-the-crc-consultation-paper/ [URL Memento: Wayback Machine]

O'Dell, E. (2013, October 29). Modernising Copyright: The Report of the Copyright Review Committee #CRC13 [Blog post]. CEARTA.IE, the Irish for Rights. Retrieved 2022-01-16, from http://www.cearta.ie/2013/10/modernising-copyright-the-report-of-the-copyright-review-committee/ [URL Memento: Wayback Machine]

O'Dell, E. (2016a, August 4). The Copyright and Related Rights (Miscellaneous Provisions) Bill 2016 is announced [Blog post]. CEARTA.IE, the Irish for Rights. Retrieved 2022-05-19, from http://www.cearta.ie/2016/08/the-copyright-and-related-rights-miscellaneous-provisions-bill-2016-is-announced/ [URL Memento: Wayback Machine]

O'Dell, E. (2016b, November 18). Copyright reform and digital deposit [Blog post]. CEARTA.IE, the Irish for Rights. Retrieved 2020-08-17, from http://www.cearta.ie/2016/11/copyright-reform-and-digital-deposit/ [URL Memento: Wayback Machine]

O'Dell, E. (2017a, May 11). Legal deposit of digital publications [Blog post]. CEARTA.IE, the Irish for Rights. Retrieved 2020-08-17, from http://www.cearta.ie/2017/05/legal-deposit-of-digital-publications/ [URL Memento: Wayback Machine]

O'Dell, E. (2017b, October 2). The copyright implications of a publicly curated online archive of Oireachtas debates [Blog post]. CEARTA.IE, the Irish for Rights. Retrieved 2022-05-11, from http://www.cearta.ie/2017/10/the-copyright-implications-of-a-publicly-curated-online-archive-of-oireachtas-debates/ [URL Memento: Wayback Machine]

O'Dell, E. (2018, October 3). Digital deposit and harvesting the .ie domain [Blog post]. CEARTA.IE, the Irish for Rights. Retrieved 2022-01-19, from http://www.cearta.ie/2018/10/digital-deposit-and-harvesting-the-ie-domain/ [URL Memento: Wayback Machine]

O'Dell, E. (2019, May 17). Not archiving the .ie domain, and the death of new politics [Blog post]. CEARTA.IE, the Irish for Rights. Retrieved 2022-01-19, from http://www.cearta.ie/2019/05/not-archiving-the-ie-domain-and-the-death-of-new-politics/ [URL Memento: Wayback Machine]

O'Dell, E. (2020, May 1). Irish copyright law must enable digital deposit (corrected and updated 11 December 2020) [Blog post]. CEARTA.IE, the Irish for Rights. Retrieved 2020-12-11, from http://www.cearta.ie/2020/05/irish-copyright-law-must-enable-digital-deposit/ [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: Digital Technology, Culture and Society*, 0(0), 1–20. DOI: 10.1080/24701475.2021.1985835

Ogden, J., Halford, S., & Carr, L. (2017). Observing Web Archives: The Case for an Ethnographic Study of Web Archiving. *Proceedings of the 2017 ACM on Web Science Conference*, 299–308. DOI: 10.1145/3091478.3091506

Ogden, J., Kurzmeier, M., Clavert, F., & Currie, M. (2022). How to Design Web Archives Research. *SAGE Research Methods: Doing Research Online*. DOI: 10.4135/9781529610147

Ogden, J., & Maemura, E. (2021). 'Go fish': Conceptualising the challenges of engaging national web archives for digital research. *International Journal of Digital Humanities*, 2(1), 43–63. DOI: 10.1007/s42803-021-00032-5

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]

Olukotun, D., & Micek, P. (2016, January 26). Five years later: The internet shutdown that rocked Egypt. AccessNow. Retrieved 2022-07-20,

fromhttps://www.accessnow.org/five-years-later-the-internet-shutdown-that-rocked-egypt/. [URL Memento: Wayback Machine]

Owens, T. (2013, December 24). Protect Your Data: Storage and Geographic Location. The Signal [Webpage]. Retrieved 2021-03-18, from //blogs.loc.gov/thesignal/2013/12/protect-your-data-storage-and-geographic-location/ [URL Memento: Wayback Machine]

Padilla, T. (2023, June 14). Libraries and archives collectively steward global memory—Directly supporting education, research, and creativity in small and large communities around the world. We need to take a leadership role in artificial intelligence. (/1) [Tweet]. Twitter. Retrieved https://twitter.com/thomasgpadilla/status/1668963837574549505 [Twitter: @thomasgpadilla]

Panitch, J. M. (1996). Liberty, Equality, Posterity?: Some Archival Lessons from the Case of the French Revolution. *The American Archivist*, *59*(1), 30–47. JSTOR

Partridge, R. C. B. (1938). The History of the Legal Deposit of Books Throughout the British Empire: A Thesis Approved for the Honours Diploma of the Library Association. London: The Library Association. [Internet Archive Open Library]

Paßmann, J., Helmond, A., & Jansma, R. (2022). From healthy communities to toxic debates: Disqus' changing ideas about comment moderation. *Internet Histories: Digital Technology, Culture and Society*, 0(0), 1–21. DOI: 10.1080/24701475.2022.2105123

Padilla, T. (2023, June 14). Libraries and archives collectively steward global memory—Directly supporting education, research, and creativity in small and large communities around the world. We need to take a leadership role in artificial intelligence. (/1) [Tweet]. Twitter @thomasgpadilla, Retrieved from https://twitter.com/thomasgpadilla/status/1668963837574549505.

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

Pennock, M. (2013). *Web-Archiving*. 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]

Pettenati, C. (2002). Electronic publishing at the end of 2001. In M. Barone, E. Borchi, J. Huston, C Leroy, P. G. Rancoita, P. Riboni, & R. Ruchti (Eds.), *Astroparticle, Particle, Space Physics, Radiation Interaction, Detectors and Medical Physics Applications: Volume 1* (Vol. 1, pp. 525–533). World Scientific. https://doi.org/10.1142/9789812776464_0076 . Preprint: https://web.archive.org/web/20170808230408/http://villaolmo.mib.infn.it/Manuscripts/10_generalities/pettenati.pdf.

Pierson, C. (1996). *The Modern State*. London: Routledge.

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

Public Record Office of Northern Ireland. (2007). Public Record Office: A Brief History [Document]. Northern Ireland: Public Record Office of Northern Ireland. Retrieved 2016-09-02, from https://www.nidirect.gov.uk/sites/default/files/publications/general-information-series-the-public-record-office-a-brief-history-2.pdf. [URL Memento: Wayback Machine]

Public Record Office of Northern Ireland. (2008). Public Record Office of Northern Ireland (PRONI): Past, Present and Future [Document]. Northern Ireland: Public Record Office of Northern Ireland. Retrieved from https://web.archive.org/web/20081120000333/http://www.proni.gov.uk/history_of_proni_-_past__present_and_future.pdf. [Web archive: Wayback Machine; source URL: http://www.proni.gov.uk/history_of_proni_-_past__present_and_future.pdf; Timestamp: 2008-11-20 00:03:33]

Public Record Office of Northern Ireland. (2016, August 19). PRONI Takedown Policy [Document]. Northern Ireland: Public Record Office of Northern Ireland. Retrieved 2020-10-06, from https://www.nidirect.gov.uk/sites/default/files/publications/PRONI%20Takedown%20Policy_0.pdf [URL Memento: Wayback Machine]

Public Record Office of Northern Ireland. (2018, November). PRONI Web Archiving Strategy [Document]. Northern Ireland: Public Record Office of Northern Ireland. Retrieved 2020-07-30, from https://www.nidirect.gov.uk/sites/default/files/publications/Web%20Archiving%20Strategy_1.pdf [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]

Raffal, H. (2018). Tracing the online development of the Ministry of Defence and Armed Forces through the UK web archive. *Internet Histories: Digital Technology, Culture and Society*, 2(1–2), 156–178. DOI: 10.1080/24701475.2018.1456739

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]

Record Nations. (2015, April 16). The Difference Between Documents & Records. Record Nations. Retrieved 2023-06-02, from https://www.recordnations.com/blog/whats-the-difference-between-documents-and-records/. [URL Memento: Wayback Machine]

Regan, J. M. (2016). Kindling the Singing Flame: The Destruction of the Public Record Office (30 June 1922) as a Historical Problem. The Old Athlone Society, pp. 107–123. Retrieved 2022-10-31, from https://www.academia.edu/30977460/Kindling_the_Singing_Flame_The_Destruction_of_the_Public_Record_Office_30_June_1922_as_a_Historical_Problem [URL Memento: Wayback Machine]

RESAW. (n.d.). About 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]

Reyes Ayala, B. (2013). Web Archiving Bibliography 2013. University of North Texas Libraries. Retrieved 2020-07-30, from https://digital.library.unt.edu/ark:/67531/metadc172362/. [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]

Rogers, R. (2013). *Digital Methods*. Cambridge, MA, USA; London, UK: MIT Press.

Rogers, R. (2017). Doing Web history with the Internet Archive: Screencast documentaries. *Internet Histories: Digital Technology, Culture and Society*, 1(1–2), 160–172. DOI: 10.1080/24701475.2017.1307542

Rogers, R. (2019). Periodising Web Archiving. In N. Brügger & I. Milligan (Eds.), The SAGE Handbook of Web History (pp. 42–56). London: SAGE Publications.

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 [Blog site]. Web90 project. Retrieved 2022-07-09, from https://web90.hypotheses.org/. [URL Memento: Wayback Machine]

Ruane, J., & Todd, J. (2004). The Roots of Intense Ethnic Conflict may not in fact be Ethnic: Categories, Communities and Path Dependence. *European Journal of Sociology / Archives Européennes de Sociologie*, *45*(2), 209–232. DOI: 10.1017/S0003975604001432

Ruest, N. (2016, June 18). Web Archives for Historical Research Group [Community group]. Zenodo. Retrieved 2022-09-14, from

https://zenodo.org/communities/wahr/?page=1&size=20. [URL Memento: Archive.today]

Ruest, N., Fritz, S., Deschamps, R., Lin, J., & Milligan, I. (2021). From archive to analysis: Accessing web archives at scale through a cloud-based interface. *International Journal of Digital Humanities*, 2(1), 5–24. DOI: 10.1007/s42803-020-00029-6

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. (2019). Exploring the 'French web' of the 1990s. In N. Brügger & D. Laursen (Eds.), *The Historical Web and Digital Humanities: The Case of National Web Domains* (pp. 145–160). London & New York: Routledge.

Schafer, V. (2020). From Print to Digital, from Document to Data: Digitalisation at the Publications Office of the European Union. *Open Information Science*, 4(1), 203–216. DOI: 10.1515/opis-2020-0015

Schafer, V., Musiani, F., & Borelli, M. (2016). Web archiving, governance and STS. *French Journal for Media Research*, no. 6/2016(6), 23. http://frenchjournalformediaresearch.com/lodel-1.0/main/index.php?id=952 [URL Memento: Wayback Machine]

Schafer, V., Truc, G., Badouard, R., Castex, L., & Musiani, F. (2019). Paris and Nice terrorist attacks: Exploring Twitter and web archives. *Media, War & Conflict*, 12(2), 153–170. DOI: 10.1177/1750635219839382

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]

Sellitto, C. (2004). A study of missing Web-cites in scholarly articles: Towards an evaluation framework. *Journal of Information Science*, 30(6), 484–495. DOI: 10.1177/0165551504047822

Sellitto, C. (2005). The impact of impermanent Web-located citations: A study of 123 scholarly conference publications. *Journal of the American Society for Information Science and Technology*, 56(7), 695–703. DOI: 10.1002/asi.20159

Senior, P. A., & Bhopal, R. (1994). Ethnicity as a variable in epidemiological research. *BMJ*, *309*(6950), 327–330. DOI: 10.1136/bmj.309.6950.327

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]

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]

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]

Song, S., & JaJa, J. (2008). Archiving Temporal Web Information: Organization of Web Contents for Fast Access and Compact Storage (Technical Report UMIACS-TR-2008-08). Department of Electrical and Computer Engineering Institute for Advanced Computer Studies, University of Maryland. Retrieved 2020-10-28, from https://drum.lib.umd.edu/handle/1903/7569. [URL Memento: Wayback Machine]

Smith, A.D. (1993). The Ethnic Sources of Nationalism. In M. E. Brown (Ed.), *Ethnic Conflict and International Security*. New Jersey, Sussex: Princeton University Press.

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

Stanford Libraries. (n.d.). Archivability [Web page]. Stanford University. Retrieved 2022-05-02, from https://library.stanford.edu/projects/web-archiving/archivability [URL Memento: Wayback Machine]

Stangor, C. (2004). *Social Groups in Action and Interaction*. New York & Hove: Psychology Press.

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]

Stember, M. (1991). Advancing the social sciences through the interdisciplinary enterprise. The Social Science Journal, 28(1), 1–14. DOI: 10.1016/0362-3319(91)90040-B

Sterne, J. (2015+). How the internet came to Ireland. TechArchives; TechArchives Digital Repository. Retrieved 2019-01-13, from https://techarchives.irish/how-the-internet-came-to-ireland-1987-97/. [URL Memento: Wayback Machine]

Stirling, P., Chevallier, P., & Illien, G. (2012). Web Archives for Researchers: Representations, Expectations and Potential Uses. *D-Lib Magazine*, 18(3/4). DOI: 10.1045/march2012-stirling [URL Memento: Wayback Machine]

Stangor, C. (2004). *Social Groups in Action and Interaction*. New York & Hove: Psychology Press. Google-Books-ID: wBXuR_DitI0C

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

Sumner, W. G. (1906). *Folkways*. New York: Dover Publications. [Internet Archive]

Sweeney, S. (2008). The Ambiguous Origins of the Archival Principle of "Provenance". *Libraries & the Cultural Record*, 43(2), 193–213. DOI: 10.1353/lac.0.0017

T | U | V | W | X | Y | Z

Tajfel, H. (1970). Experiments in Intergroup Discrimination. *Scientific American*, *223*(5), 96–103. JSTOR

Tajfel, H. (1974). Social identity and intergroup behaviour. *Social Science Information*, *13*(2), 65–93. DOI: 10.1177/053901847401300204

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178. DOI: 10.1002/ejsp.2420010202

TARA. (n.d.). TARA [Web page]. Trinity College Dublin. Retrieved 2021-11-16, from http://www.tara.tcd.ie. [URL Memento: Wayback Machine]

Taylor, A. (2013, June 8). What Is Digital Publishing, And Is It Necessary? [Web page]. Author Mingle. https://web.archive.org/web/20130712072819/http://authormingle.com/ebooks/what-is-digital-publishing-and-is-it-necessary/ [Web archive: Wayback Machine; source URL: http://authormingle.com/ebooks/what-is-digital-publishing-and-is-it-necessary/; Timestamp: 2013-07-12 07:28:19]

Taylor, C. (2017a, July 20). Ireland's digital content in danger of disappearing, specialist warns. *The Irish Times* [online]. Retrieved 2017-09-22, from https://www.irishtimes.com/business/technology/ireland-s-digital-content-in-danger-of-disappearing-specialist-warns-1.3157792. [URL Memento: Wayback Machine]

Taylor, N. (2017b). Understanding Legal Use Cases for Web Archives [Conference presentation]. International Internet Preservation Coalition General Assembly and Web Archiving Conference, London, June 2016. Retrieved 2021-12-05, from https://nullhandle.org/pdf/2017-06-16_understanding_legal_use_cases_for_web_archives.pdf. [URL Memento: Wayback Machine]

Taylor, N. (2011, June 16). Web and Twitter Archiving at the Library of Congress [Presentation slides]. Web Archive Globalization Workshop June 16, 2011, Library of Congress, USA. Retrieved 2019-06-06, from http://eventsarchive.org/sites/default/files/webandtwitterarchivingatthelibraryofcongress-110614202159-phpapp01.pdf. [URL Memento: Wayback Machine]

Teszelszky, K. (2022, November 1). *KB web collection first digital heritage on Unesco Memory of the World list* [Web page]. Koninklijke Bibliotheek. Retrieved 2022-11-01, from https://www.kb.nl/en/recent/news/kb-web-collection-first-digital-heritage-unesco-memory-world-list. [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]

The Future of Media Commission. (2022). Report of the Future of Media Commission. Retrieved 2022-07-14, from https://www.gov.ie/en/publication/ccae8-report-of-the-future-of-media-commission/ [URL Memento: Wayback Machine]

The Irish News. (2022, March 14). Sinn Féin removes thousands of media statements from its website. *The Irish News*. Retrieved 2022-03-14, from https://www.irishnews.com/news/northernirelandnews/2022/03/14/news/sinn-fe-in-removes-thousands-of-media-statements-from-its-website-2613771/. [URL Memento: Wayback Machine]

The National Archives. (n.d.). The EEC and Britain's late entry [Web page/catalogue]. The National Archives. Retrieved 2022-08-08, from https://www.nationalarchives.gov.uk/cabinetpapers/themes/eec-britains-late-entry.htm [URL Memento: Wayback Machine]

The National Archives. (n.d.). The National Archives – Home [Website]. The National Archives. Retrieved 2021-10-07, from https://www.nationalarchives.gov.uk. [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]

Thompson, D. (2008). Archiving Web Resources. In S. Ross & M. Day (Eds.), DCC Digital Curation Manual. HATII, University of Glasgow; University of Edinburgh; UKOLN, University of Bath; Council for the Central Laboratory of the Research Councils; Digital Curation Centre. Retrieved 2021-06-04, from https://www.dcc.ac.uk/sites/default/files/documents/resource/curation-manual/chapters/archiving-web-resources/archiving-web-resources.pdf. [URL Memento: Wayback Machine]

Thomson, S. D. (2016). Preserving Social Media (DPC Technology Watch Reports). Digital Preservation Coalition. Retrieved 2019-06-07, from https://www.dpconline.org/docs/technology-watch-reports/1486-twr16-01/file [URL Memento: Wayback Machine]

Tough, A. G. (2009). Archives in sub-Saharan Africa half a century after independence. *Archival Science*, *9*(3), 187–201. DOI: 10.1007/s10502-009-9078-1

Tracy, M. (2000). *Racism and immigration in Ireland: a comparative analysis*. Dublin, Ireland. Department of Sociology, University of Dublin, Trinity College.

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]

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]

Turner, J. C., Oakes, P. J., Haslam, S. A., & McGarty, C. (1994). Self and Collective: Cognition and Social Context. *Personality and Social Psychology Bulletin*, *20*(5), 454–463. DOI: 10.1177/0146167294205002

UCD Library. (n.d.). LibGuides: Irish Poetry Reading Archive: Introduction & News [Web page]. UCD Library. Retrieved 2022-10-31, from https://libguides.ucd.ie/ipra/news [URL Memento: Archive.today]

UK Web Archive. (2009, archived). F.A.Q (archived) [Archived web page]. UK Web Archive. https://www.webarchive.org.uk/wayback/en/archive/20091130210036/http://www.w

ebarchive.org.uk/ukwa/info/faq [Web archive: UK Web Archive; source URL: http://www.webarchive.org.uk/ukwa/info/faq; Timestamp: 2009-11-30 21:00:36

UK Web Archive. (2020, March 2). 15 Years of the UK Web Archive—The Early Years. Retrieved 2020-03-02, from https://blogs.bl.uk/webarchive/2020/03/15-years-of-the-uk-web-archive.html [URL Memento: Wayback Machine]

UK Web Archive. (n.d.). Frequently Asked Questions [Web page]. UK Web Archive. Retrieved 2022-06-19, from https://www.webarchive.org.uk/en/ukwa/info/faq/ [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]

UNESCO. (2003). UNESCO Charter on the Preservation of the Digital Heritage. United Nations Educational, Scientific and Cultural Organization. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000229034. Document code: IFAP-2003/COUNCIL.II/4

United Nations Secretariat. (1985). *Administrative Instruction: Regulations for the control and limitation of documentation (ST/AI/189/Add.3/Rev.2)*. United Nations. https://documents-dds-ny.un.org/doc/UNDOC/GEN/NS0/000/75/IMG/NS000075.pdf?OpenElement. [URL Memento: Wayback Machine]

United Nations Secretariat. (2008). *Regulations for the control and limitation of documentation: Attribution of authorship in United Nations documents, publications and other official papers*. United Nations. https://digitallibrary.un.org/record/635480 [URL Memento: Wayback Machine]

University College Dublin. (2021, December 20). UCD appoints Dr Sandra Collins as Librarian. Retrieved 2022-10-01, from https://www.ucd.ie/newsandopinion/news/2021/december/20/ucdappointsdrsandracollinsaslibrarian/ [URL Memento: Wayback Machine]

University of Minnesota Libraries (Ed.). (2016). *Communication in the Real World*. University of Minnesota Libraries Publishing Edition, 2016. This edition adapted from a work originally produced in 2013 by a publisher who has requested that it not receive attribution. DOI: 10.24926/8668.0401.

Vène, P. M. (n.d.). L'ordonnance de Montpellier [Web page]. BnF Essentiels. Retrieved 2022-07-26, from https://essentiels.bnf.fr/fr/histoire/temps-modernes/ccf48dc0-ff3c-4432-9b57-7e838f7a6721-francois-ier-entre-pouvoir-et-image/article/cdbda9aa-961a-4f62-97e3-af045f10b571-ordonnance-montpellier. [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]

Walsham, A. (2016). The Social History of the Archive: Record-Keeping in Early Modern Europe. *Past & Present*, *230*(suppl_11), 9–48. DOI: 10.1093/pastj/gtw033

Waniata, R. (2018, February 7). The Life and Times of the Late, Great CD: Remembering the rise (and final fall) of the late, great Compact Disc. *Digital Trends*. Retrieved 2022-06-19, from https://www.digitaltrends.com/music/the-history-of-the-cds-rise-and-fall/. [URL Memento: Wayback Machine]

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

WARCnet. (n.d.). WARCnet Meetings [Web page]. Aarhus University. Retrieved 2022-01-01, from https://cc.au.dk/en/warcnet/meetings. [URL Memento: Wayback Machine]

WARCnet. (n.d.). London Spring 2022 [Web page]. Aarhus University. Retrieved 2022-08-04, from https://cc.au.dk/en/warcnet/meetings/london-2022. [URL Memento: archive.today]

WARCnet. (n.d.). Working Groups [Web page]. Aarhus University. Retrieved 2022-05-07, from https://cc.au.dk/en/warcnet/working-groups. [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]

Waters, D., & Garrett, J. (1996). Preserving Digital Information. Report of the Task Force on Archiving of Digital Information. The Commission on Preservation and Access and The Research Libraries Group. Retrieved 2021-05-07, from https://www.clir.org/pubs/reports/pub63/. [URL Memento: Wayback Machine]

Weber, M. (2019). Browser and Browser Wars. In N. Brügger & I. Milligan (Eds.), *The SAGE Handbook of Web History* (pp. 270–298).London: SAGE Publications.

Weber, M. S., & Napoli, P. M. (2018). Journalism History, Web Archives, and New Methods for Understanding the Evolution of Digital Journalism. *Digital Journalism*, 6(9), 1186–1205. DOI: 10.1080/21670811.2018.1510293

Weber, M. (1978). Economy and Society: An Outline of Interpretive Sociology (G. Roth & C. Wittick, Eds.). Berkeley, Los Angeles, London: University of California Press.

Webster, P. (2017a). Religious discourse in the archived web: Rowan Williams, Archbishop of Canterbury, and the sharia law controversy of 2008. In N. Brügger & R. Schroeder

(Eds.), 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]

Webster, P. (2017b). 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. (2012, February 2). The AADDA project [Education; Research; Libraries; Academic blog]. AADDA Blog. Retrieved 2020-10-20, from http://domaindarkarchive.blogspot.com/2012/02/. [URL Memento: Wayback Machine]

Webster, P. (2019). Understanding the limitations of the ccTLD as a proxy for the national web: lessons from cross-border religion in the northern Irish web sphere. In N. Brügger & D. Laursen (Eds.), The Historical Web and Digital Humanities: The Case of National Web Domains (pp. 110–123). London & New York: Routledge.

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]

Weigle, M. C. (2018, September 19). On the Importance of Web Archiving. Items: Insights from the Social Sciences. Retrieved 2021-03-10, from https://items.ssrc.org/parameters/on-the-importance-of-web-archiving/

Weisbard, P. H. (2011). Oldies but Goodies: Archiving Web- Based Information. Feminist Collections: A Quarterly of Women's Studies Resources, 32(2), 14–20. Gale Document Number: GALE|A339529981.

Weiss, R. (2003, November 24). On the Web, Research Work Proves Ephemeral. *Washington Post* [online] Retrieved 2018-04-29, from https://www.washingtonpost.com/archive/politics/2003/11/24/on-the-web-research-work-proves-ephemeral/959c882f-9ad0-4b36-88cd-fb7411db118d/ [URL Memento: Wayback Machine]

Wellcome Trust Library. (2008, September 22). Web archiving: A feasibility study for JISC and the Wellcome Trust [Archived web page]. Wellcome Trust Library. Retrieved from https://www.webarchive.org.uk/wayback/archive/20080922220153/http://library.wellcome.ac.uk/node228.html [Web archive: UK Web Archive; source URL: http://library.wellcome.ac.uk/node228.html; Timestamp: 2008-09-22 22:01:53]

White, B. (2012, August). Guaranteeing Access to Knowledge: The Role of Libraries. *WIPO Magazine*, Retrieved 2022-03-15, from https://www.wipo.int/wipo_magazine/en/2012/04/article_0004.html#. [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. (2005+). Department of Enterprise, Trade and Employment. Wikipedia. Retrieved 2022-07-07, from https://en.wikipedia.org/wiki/Department_of_Enterprise,_Trade_and_Employment [URL Memento: Wayback Machine]

Wikipedia. (2007+). Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media. Wikipedia. Retrieved 2023-01-05, from https://en.wikipedia.org/w/index.php?title=Department_of_Tourism,_Culture,_Arts,_Gaeltacht,_Sport_and_Media&oldid=1130593170. [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]

Wikipedia. (2012+). Internet Memory Foundation. Wikipedia. Retrieved 2022-06-27, from https://en.wikipedia.org/wiki/Internet_Memory_Foundation. [URL Memento: Wayback Machine]

Wikipedia. (2002+). Society. Wikipedia. Retrieved 2022-06-27, from https://en.wikipedia.org/wiki/Parenthetical_referencing. [URL Memento: Wayback Machine]

Williams, R. (1973). *Communications*. Harmondsworth: Penguin. [Internet Archive]

Winn, S. (2015). Ethics of Access in Displaced Archives. *Provenance, Journal of the Society of Georgia Archivists*, *33*(1). Retrieved from https://digitalcommons.kennesaw.edu/provenance/vol33/iss1/5. [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. (2018). Digital History. In M. Tamm & P. Burke (Eds.), Debating New Approaches to History (pp. 277–300). London: Bloomsbury Publishing.

Winters, J. (2019). Negotiating the archives of UK web space. In N. Brügger & D. Larsen (Eds.), *The Historical Web and Digital Humanities: The Case of National Web Domains* (pp. 75–88). London & New York: Routledge.

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. (2018). Digital History. In M. Tamm & P. Burke (Eds.), *Debating New Approaches to History* (pp. 277–300). London: Bloomsbury Publishing.

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

Wood, H. (1930). The Public Records of Ireland before and after 1922. *Transactions of the Royal Historical Society*, 13, 17–49. DOI: 10.2307/3678487

Working Party on Legal Deposit. (1998). Report of the Working Party on Legal Deposit. Department for Culture, Media and Sport. Retrieved from http://web.archive.org/web/20110607054218/http://www.bl.uk/aboutus/stratpolprog/legaldep/report/. [Web archive: Wayback Machine; source URL: http://www.bl.uk/aboutus/stratpolprog/legaldep/report/; Timestamp: 2011-06-07 05:42:18]

Wren, J. D. (2008). URL decay in MEDLINE—a 4-year follow-up study. *Bioinformatics*, 24(11), 1381–1385. DOI: 10.1093/bioinformatics/btn127

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

Yale, E. (2015). The History of Archives: The State of the Discipline. *Book History*, 18, 332–359. JSTOR

Yale, E. (2016). Introduction: Consider the Archive. *Isis*, *107*(1), 74–76. JSTOR

Yeo, G. (2007). Concepts of Record (1): Evidence, Information, and Persistent Representations. *The American Archivist*, *70*(2), 315–343. JSTOR

Yeo, G. (2017). Introduction to the Series. In D. Thomas, S. Fowler, & V. Johnson (Eds.), *The Silence of the Archive* (pp. ix–xii). London: Facet Publishing.

Yusuf, K. F. (2013). The Role of Archives in National Development: National Archives of Nigeria in Perspective. *International Journal of Economic Development Research and Investment*, 4(2), 19–29. Retrieved from http://icidr.org/ijedri_vol4_no2_dec2013/The_Role_of_Archives_in_National_Development_Nigeria%20National_Archives_in_Perspective.pdf [URL Memento: Archive.today]

Zhou, K., Grover, C., Klein, M., & Tobin, R. (2015). No More 404s: Predicting Referenced Link Rot in Scholarly Articles for Pro-Active Archiving. *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, 233–236. DOI: 10.1145/2756406.2756940

Zierau, E. M.-B. O. (2019). A Persistent Web IDentifier (PWID) URN Namespace. Internet Engineering Task Force. Retrieved 2022-10-31, from,

https://datatracker.ietf.org/doc/draft-pwid-urn-specification-07. Internet Draft: draft-pwid-urn-specification-07 [URL Memento: Archive. today]

Zierau, E., Nyvang, C., & Kromann, T. H. (2016). Persistent Web References - Best Practices and New Suggestions. *Proceedings of the 13th International Conference on Digital Preservation (iPRES 2016), Bern, 3-6 October 2016*, pp. 237–246. Retrieved 2022-09-14, from https://fedora.phaidra.univie.ac.at/fedora/objects/o:502767/methods/bdef:Content/download.

Zittrain, J., Albert, K., & Lessig, L. (2014). Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations. *Legal Information Management*, 14(2), 88-99. DOI: 10.1017/S1472669614000255

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]

Arvidson, A., & Lettenström, F. (1998). The Kulturarw Project—The Swedish Royal Web Archive. The Electronic Library, 16(2), 105–108. DOI: 10.1108/eb045623

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]

British Library Research Repository. (n.d.). Datasets: UK Web Archive—British Library Dataset Collections. UKWA Open Data/British Library. https://bl.iro.bl.uk/collections/d09fbc16-7a76-49db-a45f-16a99c30ae3e?utf8=%E2%9C%93&cq=Dataset&sort=score+desc%2C+system_create_dtsi+desc&per_page=100&locale=en [BL Research Repository]

British Library Research Repository. (n.d.). Host Link Graph—JISC UK Web Domain Dataset (1996-2010). UKWA Open Data/British Library. DOI: 10.5259/UKWA.DS.2/HOST.LINKAGE/1

British Library Research Repository. (n.d.). JISC UK Web Domain Dataset Crawled URL Index. 1996—2013. CDX. UKWA Open Data/British Library. DOI: 10.5259/UKWA.DS.2/CDX/1

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

Darcy, E., Charthaigh, C. N., Lalor, M., Sinnott, J., & Shank, C. (Eds.). (2021). In Her Shoes: Stories of the Eighth Amendment (Ireland). Digital Repository of Ireland, Collections. DOI: 10.7486/DRI.wm11nd02p

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]

International Internet Preservation Consortium. (n.d.). International Internet Preservation Consortium - Archive-It [Web page]. Archive-It. https://archive-it.org/home/IIPC. [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]

Ken Web Archiving. (n.d.). Ken—A Better Way to Control Enterprise Data [Website]. Ken Web Archiving. Retrieved 2021-11-30, from https://ken-webarchiving.com. [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]

Library of Trinity College Dublin. (n.d.). Library of Trinity College Dublin [Web page]. Trinity College Dublin. Retrieved 2022-07-10, from https://www.tcd.ie/library/ [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]

MirrorWeb. (n.d.). MirrorWeb: Your unified compliance platform [Website]. MirrorWeb. Retrieved 2021-11-20, from https://www.mirrorweb.com/. [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 and University Library of Iceland. (n.d.). Landsbókasafn—The Icelandic web archive. National and University Library of Iceland. Retrieved 2022-03-02, from https://landsbokasafn.is/index.php?page=icelandic-web-archive. [URL Memento: Wayback Machine]

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]

National Library of Ireland. (n.d.). NLI Web Archive—Archive-It [Web page]. Archive-It. Retrieved 2021-03-01, from https://archive-it.org/home/nli [URL Memento: Wayback Machine]

National Library of Ireland. (n.d.). NLI Web Archive—Archive-It—Collections [Web page]. Archive-It. Retrieved 2022-07-04, from https://archive-it.org/home/nli/?show=Collections [URL Memento: Wayback Machine]

National Library of Ireland. (n.d.). NLI Web Archive—Archive-It—Sites [Web page]. Archive-It. Retrieved 2022-09-22, from https://archive-it.org/home/nli/?show=Sites [URL Memento: Wayback Machine]

National Library of Ireland. (2011). NLI Web Archive—General Election 2011. Archive-It. Retrieved 2022-09-27, from https://archive-it.org/collections/19959. [URL Memento: Wayback Machine]

National Library of Korea. (n.d.). National Library of Korea Web Resources Archive—OASIS [Web page]. National Library of Korea. Retrieved 2022-09-22, from https://www.nl.go.kr/oasis/. [URL Memento: Wayback Machine]

National Library of New Zealand. (n.d.). New Zealand Web Archive [Web page]. National Library of New Zealand. Retrieved 2022-06-01, from https://natlib.govt.nz/collections/a-z/new-zealand-web-archive. [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]

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]

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 Web Archive. (n.d.). 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.). UK Web Archive [Website]. UK Web Archive. Retrieved 2021-03-18, from https://www.webarchive.org.uk/ukwa. [URL Memento: Wayback Machine]

UK Web Archive. (n.d.c). UKWA Topics and Themes [Web page]. UK Web Archive. Retrieved 2022-06-19, from https://www.webarchive.org.uk/en/ukwa/category/ [URL Memento: Wayback Machine]

UKWA Open Data. (n.d.). JISC UK Web Domain Dataset (1996-2013). UK Web Archive/British Library. DOI: 10.5259/UKWA.DS.2/1

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

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]

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]

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 [GitHub]. 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 [GitHub]. 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]

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]

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]

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]

Google. (n.d.). Google Drive [Web page]. Google. Retrieved 2021-10-15, from https://www.google.com/drive/ [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]

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 (pp. 17–23). 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]

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. (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]

International Internet Preservation Consortium. (n.d.). OpenWayback—IIPC [Web page]. International Internet Preservation Consortium; Https://Netpreserve.Org/. Retrieved 2021-10-06, from https://netpreserve.org/web-archiving/openwayback/ [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. (n.d.). Save Page Now—Wayback Machine | Internet Archive [Web page]. Internet Archive. Retrieved 2021-02-02, from https://archive.org/web/. [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]

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]

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]

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

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. (2020, November 27). Jupyter notebooks for web archives. Retrieved 2022-09-13, from https://doi.org/10.5281/zenodo.4293189. DOI: 10.5281/zenodo.4293189

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]

SurveyMonkey. (n.d.). SurveyMonkey [Website]. SurveyMonkey. Retrieved 2019-05-15, from https://www.surveymonkey.com/welcome/sem/ [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]

Taguette. (n.d.). Taguette, the free and open-source qualitative data analysis tool [Website]. Taguette. Retrieved 2021-12-17, from https://www.taguette.org/ [URL Memento: Wayback Machine]

Talkbank. (n.d.). TalkBank. Retrieved 2022-10-05, from https://talkbank.org/ [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]

UK Web Archive, & Jackson, A. (2013). Host Link Graph—JISC UK Web Domain Dataset (1996-2010). British Library. Retrieved 2022-10-31, from https://doi.org/10.5259/UKWA.DS.2/HOST.LINKAGE/1 [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]

Zenodo. (n.d.). Zenodo—Research [Website]. Zenodo. Retrieved 2021-12-22, from https://zenodo.org/ [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]

Appendix A: WARST - 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: WARST - 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: WARST - 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

Appendix D: Awareness/Engagement Survey - Recruitment Email Example

Survey Recruitment Email for Academics

The following contains the text of the email sent to academics in nine universities. The link to the survey is no longer operable. Please note the inclusion of the paragraph in [square brackets] was added after an initial 76 emails were sent.

Dear Professor xx,
I would be most grateful if you would consider participating in this anonymous survey, and sharing this email with students, lecturers, and researchers to which you are associated, for their interest to participate also. This research is being carried out by Sharon Healy, a doctoral candidate at Maynooth University, and is supervised by Prof. Susan Schreibman. This survey is about the awareness of, and engagement with web archives and archived web content in Irish third-level academic institutions.
[A '**web archive**' is a resource that captures and preserves websites, blogs, and web pages, and provides access to view such content, long after it has disappeared from the live web. A web archive differs from a digital archive/ library in so far as a web archive only contains archived websites, blogs, and web pages.]
To participate in the survey, please click
here: https://www.surveymonkey.com/r/WebArchivesIR
Please note, it is equally important to attain participation from respondents who are **not aware** of, or **do not engage** with web archives, as it is to attain responses from occasional or regular users.
It should take 8-10 minutes to complete this survey. Completion of the survey is voluntary, and participants can withdraw at any time.
The survey is targeted at the following audience:

- Undergraduate students, Postgraduate students
- PhD candidates/students, Postdoctoral associates, researchers or fellows
- Senior Lecturers/Associate Lecturers, Professors/Associate Professors
- Employed researchers in a third-level educational setting or project.

Purpose of the Study
For more than two decades, national libraries and cultural heritage organisations have been archiving websites (including blogs), which are then made accessible for current and future research, long after the original website has gone or been changed. However, to date, little is known about the awareness of, or engagement with web archives, and archived web content in Ireland. Therefore, in the context of Irish third-level academic institutions, the aim of this survey is to:

- Investigate the awareness of web archives and archived websites as a resource for study/research

- Generate a better understanding of how and why archived websites are used or not used for study/research
- Explore the challenges and opportunities for using archived websites as a resource for study/research.

Confidentiality
This study is being conducted according to Maynooth University Ethics Committee guidelines and has received their approval. Your confidentiality will be kept at all times. If you have any concerns or would like any further information about this research study, please contact the researcher, sharon.healy@mu.ie or the supervisor of this research, susan.schreibman@mu.ie.
Yours sincerely
Sharon Healy
PhD Candidate in Digital Humanities
GOIPG, Irish Research Council Scholar
Maynooth University

Appendix E: Awareness/Engagement Survey - Informed Consent

Informed Consent: Awareness of and engagement with web archives, in Irish third-level academic institutions

The following is the text of the informed consent, which introduced the survey on the SurveyMonkey platform.

Awareness of and engagement with web archives, in Irish third-level academic institutions

Information*:* Thank you for taking the time to consider participating in this study. This study is being carried out by Sharon Healy, a doctoral candidate at Maynooth University, and is supervised by Prof. Susan Schreibman. It consists of an anonymous survey and is entirely voluntary. It will take approximately 8-10 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 click Next at the bottom of this page, complete the survey and press submit, and your responses will be recorded as anonymous. If you decide to participate, it is important that you fully understand what is required.

Purpose of the Study: For more than two decades, national libraries and cultural heritage organisations have been archiving websites (inclusive of blogs), which are made accessible for current and future research, long after the original website has gone or been changed. However, to date, little is known about the awareness of, or engagement with web archives, and web archived content in Ireland. Therefore, the aim of this survey is to:

- Investigate the awareness of web archives and archived websites as a resource for study/ research
- Generate a better understanding of how and why archived websites are used or not used for study/ research
- Explore the challenges and opportunities for using archived websites as a resource for study/ research.

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 first on some basic demographic information such as nationality, gender, age, role/ position. Thereafter, you will be required to answer a questionnaire on your awareness or lack of awareness of web archives, and engagement with or lack of engagement with web archives, and web archived content.

How will the information collected by this survey be used? This study is being conducted according to Maynooth University Ethics Committee guidelines and has received their approval. Your confidentiality will be kept at all times. All opinions and data will be reported in an aggregated form so that individuals will not be identified. Summaries of the results will be included as part of a PhD dissertation and in other publications associated with this research.

What if there is a problem? If you have any concerns or would like any further information about this research study, please contact the researcher, sharon.healy@mu.ie or the supervisor of this research, susan.schreibman@mu.ie

Who will have access to this data? This data will not be shared with a third party and will only be processed in a manner compatible with the purposes of this research. Sharon Healy will act as the data controller for all responses gathered and will endeavour to store this data for a period of ten years as outlined in the Maynooth University Research Integrity Policy, after which it will be destroyed. (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.)

Informed Consent: By clicking Next and submitting this survey, you are also confirming that:

- you are 18 years of age or over
- you have been sufficiently informed about the project
- you are taking part in this research study voluntarily
- you agree to have your responses stored and processed in a manner compatible with the purposes of this research.

Next

Appendix F: Awareness/Engagement Survey - Questions

Survey Questions: Awareness of and engagement with web archives, in Irish third-level academic institutions

About You

These questions allow for the exploration of any trends from the rest of the survey across nationality, age, gender, area of study/research, use of digital research resources.

Q.1 – What is your nationality?

Dropdown Box

Country Index

Q.2 – Please select your age?

Dropdown Box

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 Box

Male Female Other Prefer not to say

Q.4 – Which of the following best describes your current student/ academic/ research position in a third-level academic institution?

Check Box

- Undergraduate student
- Postgraduate student
- PhD candidate/student
- Postdoctoral associate, researcher or fellow
- Employed researcher in a third-level educational setting or project
- Senior Lecturer or Associate Lecturer
- Professor or Associate Professor
- Other (please describe)

Q.5 – Which of the following academic disciplines best describes your primary area of study/ research?

Multiple-Choice Box

- Architecture
- Arts (visual, performance, music)
- Business, Economics, Finance

- Computer Science
- Digital Arts, Digital Humanities, Digital Cultural Heritage
- Educational Science
- Engineering Science
- Geography (cartography, hydrology, meteorology, environment)
- Government / Public Administration
- Heritage and Archival Studies
- Humanities (history, archaeology, languages, literature, philosophy, theology)
- Internet Studies
- Law (criminal, civil, common, statute)
- Library and Information Sciences
- Mathematics
- Media/Communications
- Natural Sciences (biology, chemistry, physics, earth sciences, space sciences)
- Political Science
- Social Sciences (anthropology, human geography, linguistics, sociology, psychology)
- Sport and Leisure
- Other (please specify)

Q.6 – From the list below, please indicate the frequency to which you access/ use the following online/ digital resources to assist with your studies/ research?

	I ALWAYS access/use this resource for my studies/research	I SOMETIMES access/use this resource for my studies/research	I RARELY access/use this resource for my studies/research	I NEVER access/use this resource for my studies/research
World Wide Web	X	X	X	X
Digital Archives	X	X	X	X
Digital Libraries	X	X	X	X
Virtual Research Environments	X	X	X	X

Student /Academic/ Researcher Awareness of Web Archives

For more than two decades, national libraries and cultural heritage organisations have been archiving websites (inclusive of blogs), which are made accessible for current and future research, long after the original website has gone or been changed.

Web archiving entails the processes of selecting, capturing, storing, and preserving websites and web pages, and subsequently, ensuring the provision of access to such content for future research and analysis. A web archive then is a resource that stores and preserves captured websites and web content, as well as an access point to view and reference such content.

There are two types of web archives for access:

1. An online public web archive whereby access is available to the general public via the web/internet from any location.

2. A dark web archive which is only accessible onsite in a designated reading room or Library via an onsite portal.

Q.7 – Prior to commencing this survey, were you aware that the National Library of Ireland archives websites and blogs which are made accessible through the NLI Web Archive – an online public web archive? (https://archive-it.org/home/nli)

Check Box

Yes: I was aware / No: I was not aware

Q.8 – Prior to commencing this survey, were you aware that the National Library of Ireland archived the Irish domain (.ie) in 2007 and 2017 and will soon make it available as a dark archive – only accessible onsite in a designated reading room at the National Library of Ireland?

Check Box

Yes: I was aware / No: I was not aware

Q.9 – From the list of online public web archive resources below, please indicate your awareness of their existence prior to commencing this survey

Multiple-Choice Box	Yes: I was aware	No: I was not aware
Internet Archive, Wayback Machine http://archive.org/web/	x	x
PRONI Web Archive (Public Records Office of Northern Ireland) https://www.nidirect.gov.uk/services/search-proni-web-archive	x	x
UK Web Archive (British Library) https://www.webarchive.org.uk/ukwa/	x	x
UK Government Web Archive (UK National Archives) http://www.nationalarchives.gov.uk/webarchive/	x	x
UK Parliament Web Archive http://webarchive.parliament.uk/	x	x
US Library of Congress Web Archive https://www.loc.gov/websites/collections/	x	x

Q.10 – Are there any other web archives that you are aware of, that are not listed above?

Multiple-Choice Box

Yes / No

Q.11 – If you answered YES to question Q11 above, would you please enter the names of any other web archive(s) you are aware of (please use commas to separate multiple entries)

Text Box

Q.12 – To the best of your knowledge, have you ever accessed or used an online public web archive for your personal interest?

Check Box

Yes / No / Unsure

Q.13 – To the best of your knowledge, have you ever accessed or used an online public web archive, or dark web archive to assist with your studies or research?

Check Box

Yes / No

- If answer is 'NO' to Q.14 above - Go to non-user respondents
- If answer is 'YES' to Q.14 above - Go to user respondents directs to

Non-User Respondents

You previously indicated that you have you have NOT accessed or used an online public web archive, or dark web archive to assist with your studies/ research.

This section now looks at some reasons why you do not use a web archive, and whether you might access or use a web archive in the future.

Q.14 – What are your main reasons to date for not using a web archive for your studies/ research (please tick all that apply)

- I was not aware of the availability of web archives as resources for my studies/ research
- I do not know how to use a web archive for my studies/ research
- I feel that I do not have the technical skills to use a web archive for my studies/ research
- I do not know how to find archived websites relevant to my studies/ research in a web archive
- I do not know how to cite/reference an archived website from a web archive to include in my studies/research
- I am unsure of the credibility or authority of using archived websites as a primary source my studies/research
- I am unsure about copyright implications for using archived web content for my studies/research
- Other reason(s) for not using a web archive for your studies/research (please specify)

Q.15 – What is the likelihood that you will access or use the NLI Web Archive in the future for your studies/research? (National Library of Ireland online public web archive - https://archive-it.org/home/nli)

Multiple-Choice Box

Definitely Likely	Fairly Likely	Not Very Likely	Definitively Not Likely	Unsure

Q.16 – From the list of online web archive resources below, what is the likelihood that you will access or use them in the future for your studies/research?

	Definitely Likely	Fairly Likely	Not Very Likely	Definitely Not Likely	Unsure
Internet Archive, Wayback Machine	X	X	X	X	X

PRONI Web Archive (Public Records Office of Northern Ireland)	X	X	X	X	X
UK Web Archive (British Library)	X	X	X	X	X
UK Government Web Archive (UK National Archives)	X	X	X	X	X
UK Parliament Web Archive	X	X	X	X	X
US Library of Congress Web Archive	X	X	X	X	X

Q.17 – What is the likelihood that you will ever access or use a dark web archive in the future for your studies/research? (only accessible onsite in a reading room or a library via an onsite portal)

Multiple-Choice Box

Definitely Likely	Fairly Likely	Not Very Likely	Definitively Not Likely	Unsure

User Respondent

You previously indicated that you have accessed or used an online public web archive, or dark web archive to assist with your studies/ research. This section now looks at your engagement with web archives.

Q.18 – In the context of using a web archive in general, have you accessed or used a web archive for any of the following reasons (please tick all that apply)

- Personal interests
- Historical interests
- Evidential interests
- Research interests
- Cultural interests
- Technical interests
- Design/artistic interests
- Other (please specify)

Q.19 – In the context of using archived web content (archived websites, blogs, web pages), have you used archived web content from a web archive, for any of the following reasons (please tick all that apply)

- I have used archived web content as a primary source in an academic essay/assignment for my course
- I have used archived web content to document the history of an organisation in an academic essay/assignment for my course
- I have used archived web content as a primary source in a professional research report
- I have used archived web content as a primary source in a professional publication
- I have used archived web content to document the history of an organisation in a professional report/publication

- I have used archived web content as part of my teaching materials for undergraduate students
- I have used archived web content as part of my teaching materials for postgraduate students
- I have used large volumes of archived web content for content analysis/textual analysis/ discourse analysis
- I have used large volumes of archived web content for data mining/ topic modelling/ data visualisation
- I have used large volumes of archived web content for network analysis/ geo-spatial analysis
- I have used archived web content for other reasons not listed above - please specify

Q.20 – Have you ever accessed or used the NLI Web Archive for your studies/ research? (National Library of Ireland online public web archive - https://archive-it.org/home/nli)

Multiple-Choice Box

Yes / No

Q.21 – Have you ever accessed or used the following online public web archives for your studies/research?

Multiple-Choice Box

	Yes	No
Internet Archive, Wayback Machine	X	X
PRONI Web Archive (Public Records Office of Northern Ireland)	X	X
UK Web Archive (British Library)	X	X
UK Government Web Archive (UK National Archives)	X	X
UK Parliament Web Archive	X	X
US Library of Congress Web Archive	X	X

Q.22 – Are there any other online web archives that you access or use for your studies/research that is not listed above?

Multiple-Choice Box

Yes / No

Q.23 – If you answered Yes to Q23 above, would you please write down the name(s) of any other online web archive(s) you have accessed or used for your studies/research (please use commas to separate multiple entries)

Text Box

Q.24 - Have you ever accessed or used a dark web archive for your studies/research? (a dark web archive is only accessible onsite in a designated reading room or Library via an onsite portal)

Multiple-Choice Box

Yes / No

Q.25 - If you answered Yes to Q25 above, would you please write down the name(s) of any dark web archive(s) you have accessed or used for your studies/ research (please use commas to separate multiple entries)

Text Box

Q.26 - In your opinion, what is the likelihood that you will access or use a dark web archive in the future for your studies/research?

Multiple-Choice Box

Definitely Likely	Fairly Likely	Not Very Likely	Definitively Not Likely	Unsure

Final section

Q.27 – In your opinion how important is it to archive websites and blogs for current and future research, based on the following values?

Multiple-Choice Box

	Very important	Fairly important	Slightly Important	Not important	No opinion
Historical value	X	X	X	X	X
Evidential value	X	X	X	X	X
Research value	X	X	X	X	X
Cultural value	X	X	X	X	X
Technical value	X	X	X	X	X
Design/artistic value	X	X	X	X	X

Q.28 – In your opinion, how important is it to archive websites and blogs based on the following topics?

Multiple-Choice Box

	Very important	Fairly important	Slightly important	Not important	No Opinion
Indirect Government (websites of agencies deployed by the Irish Government to undertake a task, e.g. Irish Water, Nama)	X	X	X	X	X
Politics (websites/blogs of political parties, political commentators)	X	X	X	X	X
Community Groups/ Activists (websites/blogs of clubs, societies, advocacy groups, human rights groups)	X	X	X	X	X

	Very important	Fairly important	Slightly important	Not important	No Opinion
Events (websites/blogs for natural disasters, sporting events, commemoration events)	X	X	X	X	X
Election Campaigns (websites/blogs of candidates, election judicators, commentators)	X	X	X	X	X
Referendum Campaigns (websites/blogs of interest groups, referendum judicators, commentators)	X	X	X	X	X
Environment (websites/blogs which report on climate change, pollution, conservation)	X	X	X	X	X
Science (websites/blogs which report on advances in medicine, chemistry, physics)	X	X	X	X	X

Q.29 – In your opinion, do you think web archives will become important as a resource for current, medium or long-term future research in your field?

Multiple-Choice Box

	Yes	No	Maybe
Current research (next 5 years)	X	X	X
Medium-term research (5-15 years)	X	X	X
Long-term research (15+ years)	X	X	X

Q.30 - OPTIONAL: In your opinion, what will be the main challenges to use archived web content for studies/research in your field in the future?

Text Box

SUBMIT SURVEY

'Your Response has been submitted" Landing Page

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 other Irish academic institutions.
www.surveymonkey.com. 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 the researcher, 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 researcher, sharon.healy@mu.ie or the supervisor of this research, susan.schreibman@mu.ie, Maynooth University, Co. Kildare, Ireland

Please Note: 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.*

Appendix G: Awareness/Engagement Survey - Use of online NLI web archives for studies or research

Table G.1: Breakdown for position and discipline categories of respondents who indicated that they use the online public NLI Web Archive for their studies/research (=23)

Uses of the public NLI Web Archive (=23)	Undergrad	Postgrad	PhD student	Postdoc. researcher or fellow	Employed researcher	Senior Lecturer or Assoc.	Professor or Assoc.	Totals
Architecture		=1						(=1)
Educational Science					=1	=1		(=2)
Geography	=1						=1	(=2)
Humanities	=1		=2	=1	=2	=2	=2	(=10)
Law	=1	=1						(=2)
Media/Communications			=1				=1	(=2)
Natural Sciences							=1	(=1)
Nursing, Midwifery	=1							(=1)
Social Sciences			=1				=1	(=2)
Total	(=4)	(=2)	(=4)	(=1)	(=3)	(=3)	(=6)	(=23)

Appendix H: Awareness/Engagement Survey - Disciplines for respondents who indicated 'Yes' on the importance of web archives

Table H.1: Discipline categories for respondents (N=239) who indicated 'Yes' on the importance of web archives for current, medium, or long-term future

Discipline Categories	'YES' Current Research	'YES' Medium-Term Research	'YES' Long-Term Research	Total number of respondents per discipline category
Architecture	=1	=2	=2	n=2
Arts	=1	=3	=3	n=3
Business, Economics, Finance	=2	=3	=3	n=6
Built Environment	=0	=0	=0	n=1
Computer Science	=2	=5	=6	n=11
Construction Management	=0	=0	=0	n=1
Dental Science	=1	=1	=1	n=1
Digital Arts/Humanities/Heritage	=2	=3	=3	n=4
Educational Science	=5	=9	=9	n=13
Engineering Science	=6	=11	=13	n=24
Geography	=1	=3	=3	n=4
Government/Public Administration	=1	=0	=0	n=1
Health Studies/Sciences	=5	=6	=7	n=11
Heritage Studies, Archival Studies	=0	=1	=1	n=1
Humanities	=27	=35	=37	n=50
Law	=8	=9	=10	n=18
Mathematics	=1	=1	=1	n=4
Medicine, Biomedical Engineering	=1	=1	=1	n=2
Media/Communications	=2	=4	=4	n=4
Natural Sciences	=5	=10	=12	n=29
Nursing, Midwifery	=6	=6	=5	n=9
Political Science	=3	=5	=5	n=6
Psychotherapy	=0	=0	=0	n=1
Social Sciences	=18	=24	=22	n=33
Totals	(=98)	(=142)	(=148)	(N=239)