Research Data Management and Sharing Practices in the Digital Humanities with a Focus on Publisher Support: A Case Study in the Field of Web Archive Studies

Victoria Zea Truter

Student number: TRTVIC001

A minor dissertation submitted in *partial fulfillment* of the requirements for the award of the degree of Master of Philosophy in Digital Curation

Faculty of the Humanities

University of Cape Town

2021

COMPULSORY DECLARATION

This work has not been previously submitted in whole, or in part, for the award of any degree. It is my own work. Each significant contribution to, and quotation in, this dissertation from the work, or works, of other people has been attributed, and has been cited and referenced.

Signed by candidate

Signature:_____ Date: 16 July 2021

Abstract

The research problem at the centre of this study is twofold. First, not enough Research Data Management studies have been conducted in either the humanities or the Digital Humanities that present a well-developed understanding of the nature of data in these fields, or the appropriate management thereof. Second, there is a critical lack of Research Data Management and data sharing support provided to researchers in these fields. While multiple stakeholders play roles in providing such support, this study focuses on the support provided to researchers by publishers.

While the overarching study investigates data management and sharing in the Digital Humanities and how publishers support these practices, the specific case concerns the field of Web Archive Studies. The case study also gathers broader insights into Digital Humanities researchers, under which WAS is classified as a specialised field.

The purpose of the study was to explore the nature of data, and current RDM and data sharing practices of Web Archive Studies researchers, with a focus on publishers' engagement with researchers and support for said practices. The aim was to uncover ways in which publishers might better support Web Archive Studies researchers in managing and sharing their data.

The case study answered the following research questions: (1) 'What kinds of data do Web Archive Studies researchers generate and work with?'; (2) 'What RDM and data sharing practices do these researchers tend to use?'; (3) 'What challenges and limitations do they encounter when collecting, managing, and sharing data?'; (4) 'How can publishers better support Web Archive Studies researchers in managing and sharing their data?'.

The study is exploratory in nature and uses a convergent mixed-methods approach based within an interpretive paradigm. Three semi-structured interviews (using predominantly open-ended questions) and a questionnaire (including predominantly multiple-choice questions) were conducted. A content analysis approach was used to analyse qualitative data, while quantitative data were interpreted using inferential statistics. The populations sampled included publishers and Web Archive Studies researchers.

The study found that Web Archive Studies researchers tend to manage their data proficiently. The biggest gaps in their current practices concern data sharing in formal repositories due to challenges like legal restrictions. Additional findings reveal a lack of funding for Research Data Management and data sharing in this field, as well as a lack of guidance and training from publishers for Web Archive Studies researchers.

Key recommendations include the following: (1) publishers should develop guidance specific to Web Archive Studies researchers' RDM and data sharing needs; (2) publishers should focus on sharing methodological processes, audit trails, and research instruments, rather than sharing data for Web Archive Studies and other humanities subjects. These actions would promote transparency in subject areas for which data sharing is often not possible due to legal restrictions, among other challenges.

Acknowledgements

I would like to thank the interviewees and questionnaire respondents who contributed to this study. This study could not have happened without their willingness and enthusiasm. I would also like to thank the following organisations who agreed to distribute the questionnaire throughout their networks: the Digital Humanities Summer Institute (DHSI), the International Internet Preservation Consortium (IIPC), the Web ARChive Studies Network (WARCnet), the Digital Research Infrastructure for the Arts and Humanities (DARIAH), Archives Unleashed, the Digital Preservation Coalition (DPC), and the Engaging with Web Archives (EWA) conference.

I wish to thank my supervisor, Michelle Kahn, for her ongoing support and advice throughout this study. I would also like to acknowledge Janet Remmington, Caroline Sutton, Emma Greenwood, and Jane Winters for their support and encouragement.

Roderick MacLeod deserves thanks for many things, but especially for cheering me on, and for always cheering me up – particularly when the end has seemed most distant. To my good friends and study group members, Julie-Anne Lothian and Jessica Barraclough, who have made the solitary process of dissertation writing a little less lonely.

I am also grateful for the financial support received from Taylor & Francis for this research project.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CESSDA – Consortium of European Social Science Data Archives

COS – Centre for Open Science

DARIAH – The Digital Research Infrastructure for the Arts and Humanities

DAS – Data Availability Statement

DCC – The Digital Curation Centre

DHSI – The Digital Humanities Summer Institute

DMP – Data Management Plan

DPC – Digital Preservation Coalition

EWA – Engaging with Web Archives conference

FAIR – Findable, Accessible, Interoperable, Reusable

GDPR – European Union's General Data Protection Regulation

IIPC – The International Internet Preservation Consortium

IP – Intellectual property

PCAST – President's Council of Advisors on Science and Technology

RDA – Research Data Alliance

RDM – Research Data Management

RDMLA – Research Data Management Librarian Academy

STM – International Association of Scientific, Technical, and Medical Publishers

TOP – Transparency and Openness Promotion Guidelines

UCT – The University of Cape Town

WARCnet – The Web ARChive Studies Network

WAS – Web Archive Studies

Referencing style

This dissertation is styled according to the University of Cape Town's Author-Date referencing style.

Chapter 1: Introduction to the study

1.1 Introduction

According to Jones, Grant and Hrynaszkiewicz (2019:1), "[o]pen science/open scholarship/open research – whichever term is preferred – refers to a set of perspectives, techniques and tools that seek to enhance the transparency, reproducibility and overall robustness of research". 'Open Research' considers "how the whole research lifecycle can be opened up, with the ultimate aim of improving the quality and integrity of research" (Jones, Grant & Hrynaszkiewicz, 2019:1). While open research is multi-faceted (Jones, Grant & Hrynaszkiewicz, 2019:1), and includes open access publishing, open-source software, open education resources, and science communication, among others, the current study focuses on open research data, which is essential to the open research agenda.

For data to be considered 'open', it needs to be managed throughout its lifecycle: "[p]roper management of data throughout the research process is crucial for making it openly accessible, intelligible, assessable and usable" (Ünal et al., 2019). This practice is referred to as Research Data Management (RDM), which has been a topic of great interest for some decades already. However, much of the research, development, and infrastructure in RDM is situated within scientific research fields. There are significantly fewer RDM-focused studies in and for the humanities due to fewer resources and collaborations, the diverse range of qualitative data types, and frequently complex data management requirements relating to policy, ethics, and ownership (Jones, Grant & Hrynaszkiewicz, 2019:4). As such, current support structures from stakeholders within the research ecosystem – publishers, institutional libraries, and funders – are more aligned to the RDM and data sharing needs of researchers in the sciences, rather than those in the humanities.

These challenges (discussed in detail in Chapter 2) are important to resolve to "gain better access to rare materials, making the representation of marginal or underrepresented positions stronger" (Digital Research Infrastructure for the Arts and Humanities [DARIAH], n.d.a). To overcome these challenges, one needs to understand the nature of humanities data and how to manage it appropriately. Researcher support is also required from various stakeholders across the research ecosystem. By developing a case study focusing on the field of Web Archive Studies (WAS)[1], the current research explores the RDM and data sharing practices of Digital Humanities researchers and highlights ways in which publishers might better support them. As Vasilevsky et al. (2017) state, there

[1] The decision was made to abbreviate 'Web Archive Studies' throughout due to its prominence within this study.

is potential for publishers to "play an important role in facilitating and enforcing data sharing". By providing greater support to Digital Humanities researchers, publishers can directly contribute to RDM and data sharing best practices in this field.

To clarify, this study considers the Digital Humanities as a discipline within the broader humanities, and reference is made to both throughout this dissertation, but only insofar as they are relevant to the study.

1.2 The emergence of Research Data Management

In order to understand why open research data and RDM have garnered increasing interest, one first needs to contemplate the sheer volume of extant data. According to Marr (2018), in May 2018, "2.5 quintillion bytes of data" were being created each day, and such volumes are increasing exponentially as time goes on. The latter has been aptly termed "the data deluge" (President's Council of Advisors on Science and Technology [PCAST], 2007:35).

The 'data deluge' phenomenon presents a curious conundrum. On one hand, there is great potential inherent in a "data-rich world" (Borgman, 2012:1059), including medical, scientific, and humanistic discovery. On the other hand, such volumes of data are "overwhelming the capacity of institutions to manage it and researchers to make use of it" (PCAST, 2007:35).

In order to benefit from the data deluge, funding bodies worldwide have begun mandating that funded researchers make their research data accessible. As Sturges et al. (2015:2445) note, "[o]rganizations wish to maximize value in their investment, and there is growing belief among funders that access to data is part of that value". However, by mandating that data be shared, funders have placed extraordinary pressure on researchers themselves, who do not necessarily know how best to do so. Furthermore, according to Borgman (2012:1066), these pressures are increasingly exerted by institutions and publishers too.

As Faniel and Zimmerman explain (2011:59), in the context of the 'data deluge', the aim "to make data accessible and usable to anyone, anytime, anywhere, and for any purpose" has become "increasingly ambitious". Borgman (2012:1059) details some of the difficulties of this process:

> [i]f the rewards of the data deluge are to be reaped, then researchers who produce those data must share them, and do so in such a way that the data are interpretable and reusable by others. Underlying this simple statement are thick layers of complexity about the nature of data, research, innovation, and scholarship, incentives and rewards, economics and intellectual property, and public policy.

This is where RDM comes in, which provides best-practice structures regarding the planning, creation, organisation, preservation, and sharing of research data for the purposes of re-use. RDM "enable[s] innovation, the avoidance of duplication, reproduction with an eye toward verification, and concrete return on public investment" (Poole & Garwood, 2019:9). While policies and mandates have been enacted with the intention of maximising the benefits of sharing data, RDM emerged as a way to address the complexities of doing so. In other words, if sharing data is the end-goal, RDM theoretically provides researchers with the tools to achieve said goal, and to navigate various unavoidable challenges relating to workflow, ownership, technology, ethics, politics, and policy.

1.3 Research Data Management: a foundation in the sciences

Most of the extant work in the field of RDM is overwhelmingly situated in the sciences, and to a lesser degree in the social sciences. These fields tend to receive the most funding, and therefore experience more immediate pressure to respond to data sharing mandates. Because of this, "[s]haring research data has become standard practice in disciplines where there is a collaborative scientific culture, such as physics, astronomy, and genetics" (Gómez, Méndez & Hernández-Pérez, 2016:546). In short, foundational exercises in RDM and data sharing have been attended due largely to rigid data sharing requirements in scientific research fields. While the sciences enjoy more established RDM practices, these may still prove useful for RDM guidance in other disciplines (Beagrie, 2019:10).

1.4 Research Data Management and the humanities

An academic domain that has received considerably less attention with regard to RDM is the humanities. Indeed, humanities data are often not even strictly considered to be 'data', as many humanities scholars frequently refer to their research materials as "sources" or "documents" (DARIAH, n.d.a). Wyatt (2019) concurs, stating that "data is a poor term for capturing the richness of the materials that humanities scholars deal with". Drucker (2011) furthers these assertions, arguing that the term 'data' be re-defined as 'capta' for humanities subjects.

That the notion of data is relatively undeveloped in the humanities since such data often assumes a qualitative form "drawn from records of human culture, whether archival materials, published documents, or artifacts" (Borgman, 2012:1061). Such materials "don't easily fit the quantitative definition of 'data sets'" (Jones, Grant, and Hrynaszkiewicz, 2019:1). Humanities data are also often secondary in nature, producing challenges around ownership and reusability: because humanities researchers often do not own their data, the ways in which they can utilise it are limited (DARIAH, n.d.a). Understanding the scope of humanities data and how it behaves proves to be a complex task,

one that differs vastly from – for example – understanding quantitative datasets generated by a primary investigator (which is more common in scientific fields).

Furthermore, it should be noted that funder investment is not evenly spread across disciplines (Waters, 2007 in Poole. 2013), and there is generally less funding allocated to the humanities. Table 1 shows the number of research and review articles published between 2017 and 2019 for two distinct Journal Citation Report categories in Web of Science (Clarivate Analytics, 2020): 'Humanities Multidisciplinary' and 'Multidisciplinary Sciences'.[2] The table illustrates how many articles had a declared funder: between 2017 and 2019, 51 per cent of the science articles had a declared funder while the same was true for only 5 per cent of the humanities articles. Zuckerman and Ehrenberg (2009:124) also confirm that "financial support for the 'academic humanities' is […] scarce".

Table 1. *Number and percentages of humanities and sciences articles published from 2017-2019 with declared funders. (Web of Science, Clarivate Analytics).*

Web of Science JCR category	No. articles published (2017-2019)	No. articles with declared funder (2017-2019)	% articles with declared funder (2017-2019)
Multidisciplinary Sciences	213,018	109,689	51%
Humanities, Multidisciplinary	35,314	1,733	5%

Due to differing investment levels, more pressure is ultimately placed on researchers in the sciences, which has therefore emphasised the importance of data curation for almost two decades (Poole, 2013). In their article exploring publicly funded Digital Humanities RDM research, Poole and Garwood (2019:4) found that "funding agencies generally issue vague or inconsistent data management requirements and provide few resources" for general humanities researchers. There is therefore far less incentive for humanities researchers to incorporate detailed RDM practices into their research workflows.

One area of the humanities that *is* attracting more funding, and which does seem to have focused more on RDM, is the Digital Humanities (Poole & Garwood, 2019:1). The definitions of Digital Humanities are contested, which is discussed further in Chapter 2. For the purpose of immediate discussion, the Digital Humanities embody a research methodology that applies computational tools to a diverse range of humanities fields. Digital Humanities projects thus include methods such as textual data mining and digital mapping, among others, and are applied across various disciplines in

[2] This data was gathered from Clarivate Analytics' Web of Science database in October 2020. Web of Science is a global citation database (Web of Science Group, n.d.), which provides an indexing service and is used to track the impact of research publications, funders, and publication categories across disciplines.

the humanities, such as history, visual arts, anthropology, archaeology, media studies, and philosophy. Digital Humanities projects also often involve larger volumes of data than more traditional humanities fields. It is also quite common for such projects to involve collaborations across different disciplines and countries, therefore often involving multiple team members situated in different time zones. Therefore, as a field that uses larger volumes of data, experiments with digital methods, and which requires diverse skill-sets and collaborative teamwork, the Digital Humanities could benefit significantly from robust RDM and data sharing practices, and it is in this field that my case study is situated.

1.5 Background to the study

This study focuses on a specific subject area within the Digital Humanities, namely Web Archive Studies (WAS), which involves the archiving of the web and the use of web archive data for research. While the former involves preserving digital objects from the web (Teszelszky, 2019:14), the latter uses the information stored in a web archive to investigate a given topic – for example – from a sociological or historical perspective.

Brügger and Laursen (2019:1) note the significance of the web and its implications for society over the past three decades. For them, "the web – and the web of the past – is an essential part of our cultural heritage, and thus also an important source for scholars of the Digital Humanities, including scholars from any field of study with a historical approach" (Brügger & Laursen, 2019:1).

Although the web is and was fundamental to how society functions, websites and webpages are constantly disappearing or being replaced (Brügger & Laursen, 2019:1). To counter this loss of information, "large national web archives have been established, and in a number of countries entire national webs have been archived" and this "treasure trove is just waiting to be studied" (Brügger & Laursen, 2019:1). Indeed, Milligan and Smyth (2019:46) highlight that web archiving has already proven crucial to understanding collective histories. That said, as Schroeder & Brügger (2017:1) confirm, the web ultimately remains a largely "untapped source for research". WAS, then, is a field of study deserving of further attention and investigation.

As the field is still new, most published literature focuses on the processes and challenges involved in the practice of web archiving, providing a basis for understanding how the underlying datasets – or corpora – are established for research. None of the literature examined discusses WAS in the context of RDM specifically. However, it is clear that this field of study would benefit from RDM studies, given the complexity of the web itself, of current WAS research methods and processes, as well as the location of WAS within the Digital Humanities.

Speaking to the complexity of the web and the argument for emphasising RDM within WAS, Teszelszky (2019:14), referring to Brügger (2010), describes the web as a complex ecosystem composed of various layers and digital data types. These include: the entire world wide web; specific web spheres such as domains; individual web pages; and the smaller elements thereof like images, HTML code, and text. As Goggin and McLellan (2017:9) highlight, citing Brügger (2010), "web pages are entirely unlike traditional media such as books or films, because they are made up of hyperlinks that result in a dense 'strata' rather than a single medium".

Furthermore, the volumes of extant data on the web are immense, and growing: Winters (2019:76-77), citing Webster (2013) and Hartmann (2015), explains that the national Web Domain Crawl at the British Library captured 31 terabytes (TB) of data in its annual crawl in 2013, and 56 TB only a year later. Hockx-Yu, Laursen and Gomes (2019:64), note that some countries, in response to rapidly expanding volumes of data, have even begun to include the web in Legal Deposit legislation to ensure its "systematic collection and preservation".

Additionally, since web archives are reconstructions of what the web once looked like at a particular point in history, it can therefore never reflect the web as it actually is in the present (Teszelszky, 2019:15; Schroeder & Brügger, 2017:9). It is critical, therefore, for a web archivist to have a clear strategy as to which data are kept, and which are discarded (Schroeder & Brügger, 2017:9).

A final point regarding web archive data complexity is that when it comes to using the data preserved in web archives, a corpus (or 'set' of data) needs to be created that includes only certain data from an archive. While a web archive itself generally preserves the whole web as it looked at a particular moment in history, a corpus would usually only include data taken from a web archive that is relevant to a particular study. Brügger, Laursen and Nielsen (2019) discuss the complexities involved in establishing a corpus and emphasise the importance of keeping a clear record of the methods and processes used to create it. As Schroeder and Brügger (2017:9) state, "the ways in which things are collected, made accessible and documented have an impact on how they can later be used by researchers". If web archive data are to be shared, re-used, and/or validated, the data need to be treated and shared appropriately.

1.6 Problem statement

As has been outlined above, although there appears to be a growing body of RDM studies within the humanities, especially the Digital Humanities, there are comparably fewer than in the sciences. Due to humanities data often being qualitative and secondary (though Digital Humanities data can also be

quantitative), a different approach to RDM is required than what has already been established in scientific fields.

The research problem at the centre of my dissertation is twofold. First, not enough RDM studies have been conducted in either the humanities or the Digital Humanities that present a well-developed understanding of the nature of data in these fields, or how to manage it most appropriately. Second, there is a critical lack of support provided to researchers in these fields in managing and sharing their data. While multiple stakeholders play a role in providing such support to researchers, this study focuses on support provided by publishers specifically.

The current study aims to address these gaps by investigating data management and sharing in the Digital Humanities, and how publishers are supporting these practices. The specific case that will be investigated concerns the field of WAS. The case will contribute fresh insights into this new area, as well as the overarching discipline of Digital Humanities, situated in the humanities more broadly.

1.7 Research objectives and questions

This study will explore the nature of data and the current RDM and data sharing practices of researchers in the field of WAS. The study will also discover the way in which publishers are currently engaging with and supporting WAS researchers with their RDM and data sharing. In doing so, the study will propose ways that may assist publishers in better supporting WAS researchers in managing and sharing their data appropriately.

In line with the above objectives, the following research questions are addressed:

- RQ 1: What kinds of data do Web Archive Studies researchers generate and work with?
- RQ 2: What RDM and data sharing practices do Web Archive Studies researchers tend to use?
- RQ 3: What challenges and limitations do Web Archive Studies researchers encounter when collecting, managing, and sharing data?
- RQ 4: How can publishers better support Web Archive Studies researchers with regard to data management and data sharing?

Certain aspects about the nature of data as well as data management and sharing practices will be detailed for the humanities and the Digital Humanities more generally, but only as is relevant to Web Archive Studies' location within these disciplines.

1.8 Rationale

The decision to develop this case study was made in the hopes of clarifying ways in which publishers might more adequately support researchers in the Digital Humanities, foster sound RDM practices, and contribute collaboratively and equitably to a diverse and global open research agenda.

1.9 Overview of the methodology

This study uses a convergent mixed methods approach within an interpretive paradigm. Three interviews and a questionnaire were used to develop an exploratory case study, based in the field of WAS. A research data lifecycle was used to inform the research instruments and the data analysis.

1.10 Delimitations of the study

Although various stakeholders provide support to Digital Humanities researchers in managing and sharing their data, this study focuses on support provided by publishers. The current case study (a research design that necessarily delimits a study) focuses on WAS as it is an emerging field of interest.

1.11 Organisation of the report

The current research report is presented in five chapters. Chapter 1 presents an introduction and provides some context to the study. Chapter 2 presents literature most relevant to the study's research objectives. Chapter 3 presents the study's research approach, paradigm, design, and instruments. Chapter 4 presents the data collected during the research, and Chapter 5 discusses key findings in the context of relevant literature, and issues some concluding remarks.

1.12 Concluding remarks

RDM and data sharing are practices necessary for keeping research data as open as possible. There is a notable lack of literature on RDM and data sharing in the Digital Humanities, and a lack of stakeholder support provided to humanities researchers in this regard. In order to help fill this gap, the current study investigates the RDM and data sharing practices of WAS researchers, and the level of current support provided to them by publishers. This research aims to outline ways in which publishers might support such researchers in the future.

Chapter 2: Literature review

2.1 Introduction

RDM and the Digital Humanities are evolving fields, necessitating that scholars stay abreast of recent developments. As such, all literature discussed in this chapter was published within the last decade – between 2010 and 2020, and spans various countries including Denmark, Norway, Italy, Germany, France, the Netherlands, Ireland, Canada, the USA, the UK, and Australia. Most authors are based in Western Europe and North America – the regions in which RDM and Digital Humanities research infrastructure[3] is most developed. Given that the current study is being conducted in South Africa, reference is made to some relevant regional literature. Sources herein include monographs, academic books and journals, blog posts, news articles, web pages, and reports. Of notable import is a seminal text regarding web archives titled *The Historical Web and Digital Humanities: The Case of National Web Domains* (Brügger & Laursen, 2019).

This chapter presents literature pertaining to the current research objectives, especially that which concerns the nature of data, as well as RDM and data sharing practices in the Digital Humanities and the humanities more broadly. Further discussions around how publishers and other stakeholders currently support and engage with these practices are outlined, and emphasis is placed on the various challenges that researchers might encounter regarding RDM and data sharing. Although there seems to be no extant literature specifically investigating RDM and data sharing practices in the field of WAS, the chapter shows how published WAS research relates to these topics. A brief overview is also given of the Jisc Research Data Lifecycle (JRDL) – this model informs the study's research instruments, data analysis, and interpretation.

While the case study focuses on Web Archive Studies, the lack of RDM and data sharing related literature in this field means it is necessary to present relevant literature in the wider fields of the Digital Humanities and the traditional humanities, as Web Archive Studies is related to and situated within both fields.

2.2 Research data

Data are evidently of great value to research (Koopman, 2016:1), but present a number of challenges to researchers and RDM processes. This section shows how data are currently discussed and

[3] Almas, in her article on infrastructure in the Digital Humanities, cites Mark Parsons, the Secretary General of the RDA, as such: "infrastructure can be defined as 'the relationships, interactions and connections between people, technologies, and institutions that help data flow and be useful (Parsons 2015)'" (Parsons qtd. In Almas, 2017:1).

understood – both generally as well as in the context of humanities and Digital Humanities – by various scholars in the field.

2.2.1 An overview of research data

The Organisation for Economic Co-operation and Development (OECD) defines research data as "factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings" (OECD, 2007:13).

Beagrie (2019:8) confirms that there are multiple definitions for 'research data', noting the increased attention paid to "software, scripts, physical samples" and even annotations as types of data (Beagrie, 2019:8). Beagrie (2019:8) ultimately adopts the 2016 Concordat of Open Research Data's definition of research data as "print, digital or physical data and other evidence that can be used to validate findings". Borgman (2012:1066), however, explains that any definition of research data depends on the context in which said data are gathered, presented, or used. This complicates efforts to find a definition since it is not always obvious what data actually are – it exists in multiple forms and is often non-transferable or un-shareable (Borgman, 2012).

The Digital Curation Centre (DCC) – a hub of knowledge regarding research data and RDM – alternatively defines research data according to the International Organisation for Standardisation's (ISO) Open Archival Information System (OAIS) Reference Model published in 2005: "[a] reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing" ("Data", n.d). Examples given include "a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen".

As illustrated, the definitions of research data are varied. While there are commonalities, there is no standard definition, which presents a number of challenges for stakeholders, including researchers, librarians, repositories, and publishers. Different disciplines also understand data differently, and therefore require different approaches to data management.

2.2.2 Data in the humanities

Borgman (2012:1061) states that, compared to other disciplines, "[t]he notion of data is least well developed in the humanities". Indeed, as Henry (2014:347) notes, "[a]t the start of the 21st century,

it was uncommon to hear humanities scholars talk about their research in terms of data". Schöch (2013) concurs with the latter in an essay on the nature of data in the humanities:

> Most of my colleagues in literary and cultural studies would not necessarily speak of their objects of study as 'data'. If you ask them what it is they are studying, they would rather speak of books, paintings and movies; of drama and crime fiction, of still lives and action painting; of German expressionist movies and romantic comedy. […]. Maybe they would talk about what they are studying as texts, images, and sounds. But rarely would they consider their objects of study to be 'data'.

The concept of data is thus relatively new in humanities fields, and one must necessarily establish how it is currently understood. There are key challenges, touched on in Chapter 1, specific to humanities data that are discussed below.

First, while research data in the humanities might differ from other disciplines such as the sciences, there are varied conceptualisations of 'data' within the humanities itself (Borgman, 2008 in Gómez, Méndez & Hernández-Pérez, 2016:547). Owens (2011) asserts that humanities scholars might understand data as a constructed artifact, an authored text, or as information that can be processed by a computer (the latter pertains largely to the Digital Humanities). The lack of a standard definition for what data is in the humanities thus presents a challenge. In order to manage and share data effectively, a streamlined infrastructure (including file formats and naming conventions, Data Management Plans [DMPs], RDM training and guidelines, data sharing policies, metadata schema, and repository capacity) must be able to accommodate the various specificities of humanities data. If data types and definitions are too varied, data management becomes a complex process (Gómez, Méndez & Hernández-Pérez, 2016:547).

Second, humanities data are often (but not always) of a qualitative nature, including audio, video, text, maps, newspapers, academic journals, images, and even administrative records (Gómez, Méndez & Hernández-Pérez, 2016:547). These data types are often analogue, not machine-readable, and require contextual interpretation which can complicate the process of data analysis (Schöch, 2013).

Third, humanities researchers are more likely to use pre-existing secondary data over generating their own (Poole & Garwood, 2019:2). Gómez, Méndez & Hernández-Pérez (2016:547) locate this as a fundamental issue. Borgman (2009:20) explains that secondary data cannot always be shared due to limitations caused by intellectual property (IP) and third-party ownership. The use of secondary data also means that humanities researchers often inherently generate less new data than those in the sciences (Gómez, Méndez & Hernández-Pérez, 2016:547; Ayris, 2017:49).

In summary, humanities data are diverse and difficult to define, and are often qualitative and secondary in nature. As Poole and Garwood (2019:2) note, humanities data are "[f]ragile, fluid, and

manipulable" and show "extraordinary variegation and complexity". RDM for humanities subjects can therefore be complex, and some such complications also affect the Digital Humanities.

Ultimately, as Borgman (2009:20) notes, understanding what constitutes data in the traditional humanities, understanding data sources, and how the data are managed, shared, and reused can greatly assist the RDM and data sharing goals of Digital Humanities projects.

2.2.3 Data in the Digital Humanities

It should be noted that defining the 'Digital Humanities' is a difficult endeavour (Brügger & Laursen, 2019:1), and that current definitions are debated (Rendix & Laursen, 2014:1). According to Klein and Gold (2016), it can be difficult to know what Digital Humanities work actually involves. For Schaffner & Erway (2014:7), this is largely due to the Digital Humanities being an emerging field, as is WAS.

Schaffner and Erway (2014:7) provide a simple working definition for the Digital Humanities, borrowed from Waters (2013): the "'application of digital resources and methods to humanistic inquiry'". Frischer's (2009 in Borgman, 2009:3) definition refers to "the application of information technology as an aid to fulfill the humanities' basic tasks of preserving, reconstructing, transmitting, and interpreting the human record". Both definitions provide clarity regarding the Digital Humanities' link to the traditional humanities, but further discussion is necessary.

Cohen (2010) refers to the Digital Humanities as an 'umbrella term' covering "a wide range of activities, from online preservation and digital mapping to data mining and the use of geographic information systems". Rendix and Laursen (2014:1) confer: "[f]rom its inception […] the term 'digital humanities' has been a hypernym covering several factions and methodological and theoretical approaches". For Poole and Garwood (2019:1), the field even includes aspects such as archiving and record keeping, 3D modelling and visualisation, and gaming.

These discussions provide important context around what the Digital Humanities *are*, but do not necessarily speak to an understanding of its data. Borgman (2012:1061) states that the term 'data' is more commonly used in the Digital Humanities than in traditional humanities research, perhaps since Digital Humanities data are *digital* – either digitised or born-digital – thus naturalising the term 'data'. Additionally, Digital Humanities projects often use and/or generate much larger volumes of data than traditional humanities research projects due to greater (computerised) capacities of their systems and infrastructure to process digital data. While there are differences between qualitative and quantitative data when comparing the Digital and traditional humanities, both fields face a common challenge in the use of secondary data. Lastly, given the diversity of the data, and the interdisciplinary

and collaborative nature of the Digital Humanities, projects often require meticulous data management strategies.

2.3 Research Data Management practices and challenges

In order to maximise the benefits and uses of data, appropriate management is necessary (Ng'eno & Mutula, 2018:28). This section will present some of the recommended RDM and data sharing practices using a research data lifecycle as a guide, and outline some of the challenges (generally, and in relation to the humanities and Digital Humanities) that arise in managing and sharing research data during various stages of the lifecycle. There is little extant literature regarding RDM practices of WAS researchers, but potential RDM challenges for such researchers are still explored.

2.3.1 RDM practices as outlined in the Jisc Research Data Lifecycle

Poole (2013) states that any digital curation endeavour "depends upon a lifecycle approach" in which "all stages and actions are identified, planned, and implemented in the appropriate order". A data lifecycle includes consideration of the selection, categorisation, sharing, archiving, and disseminating of data (Poole 2013); this structured approach ensures "the maintenance of authenticity, reliability, integrity and usability of digital material" (Poole, 2013).

A lifecycle approach provides a best practice structure for managing data of any kind, and might refer to various data lifecycle models, including the Data Documentation Initiative (DDI) model (Data Documentation Initiative [DDI], n.d), the UK Data Service lifecycle (UK Data Service, n.d), and the DCC's lifecycle model (DCC, n.d.a). The current study utilises the JRDL (Jisc, n.d.b) as a framework for investigation: since it is a generic model applicable to any discipline, it fits the specificities of the Digital Humanities and the current study.

The JRDL (Jisc, n.d.b) involves six consecutive stages and is accompanied by a toolkit of resources that detail the various actions involved in each stage. Figure 1 presents the JRDL lifecycle model, and the descriptions of each stage that follow are in reference to the toolkit (Jisc, n.d.c).

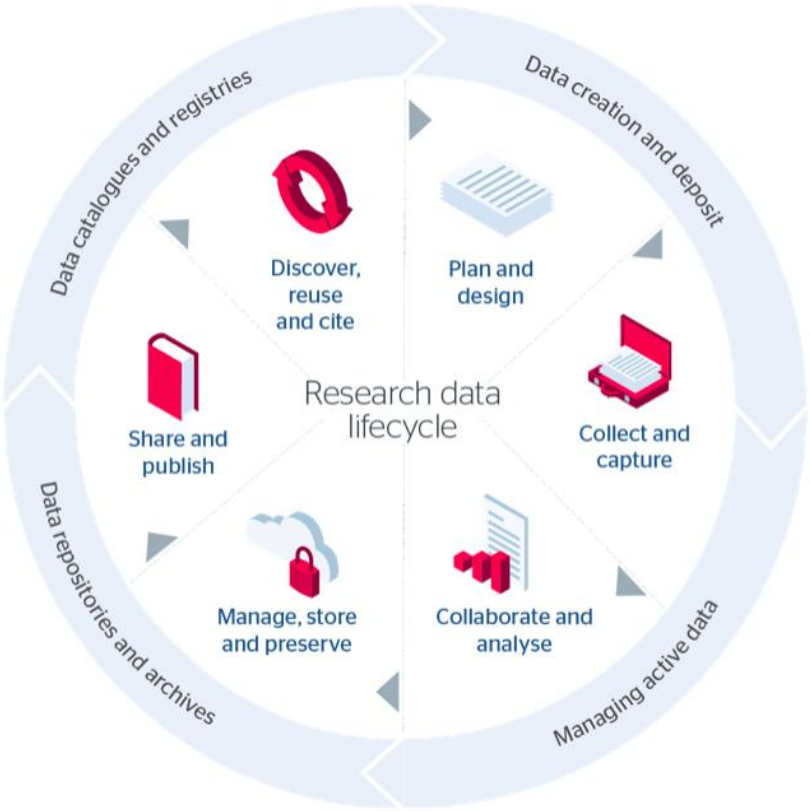

Figure 1. *Jisc Research Data Lifecycle (Jisc, n.d.b) (reproduced under a Creative Commons CC BY-ND licence).*

The 'plan and design' stage (situated within the broader 'data creation and deposit' stage) occurs at the start of a research project. It includes actions such as: finalising a DMP that outlines how one will collect, manage, and store one's data; ensuring compliance with data management and data sharing policies (institution, funder, or publisher related); estimating the costs involved in creating, processing, and managing the data (referring to computers, external hard drives, or data analysis software); and finally, using an RDM checklist to ensure no steps are forgotten during the planning phase (Jisc, n.d.b).

The 'collect and capture' stage (spanning the 'data creation and deposit' and 'managing active data' stages) include actions such as: ensuring that data are stored and backed-up; assigning metadata to ensure searchability; and managing file formats, naming conventions, and folder organisation (Jisc, n.d.b).

The 'collaborate and analyse' stage (forming part of the 'managing active data' stage) includes actions such as: installing appropriate collaboration processes; creating a library of resources for the research team; defining roles and responsibilities for each team member (for example, giving one person responsibility for maintaining the dataset, and giving others the responsibility of gathering data);

sharing data with team members; and finally, tracking and presenting data through the use of visualisations, graphs, or charts (Jisc, n.d.b).

The 'manage, store, and preserve' stage (under the umbrella stage 'data repositories and archives') includes actions such as: continually assessing and adapting DMPs; ensuring data are preserved for continued access and future re-use; ensuring the data are secured and made available only to those requiring access; and ensuring that any software generated are managed appropriately (Jisc, n.d.b).

The 'share and publish' stage (spanning the 'data repositories and archives' and 'data catalogues and registries' stages) includes actions such as: publishing data in a repository as supplementary material to a publication, or in a data journal; ensuring that the data are licenced for re-use and that IP and copyright do not limit the data's re-use; ensuring that the data are assigned a persistent identifier allowing others to discover and cite the data regardless of its location; ensuring that publication contains a Data Availability Statement (DAS) that links to the data (Jisc, n.d.b).

The 'discover, re-use, and cite' stage (within the 'data catalogues and registries' stage) refers to pre-existing data generated by third parties, and includes actions such as: finding data appropriate to a researcher's study (for example, in a repository); ensuring the data license allows for re-use; ensuring the data are cited appropriately; and sharing any new data generated using the authors' original data in accordance with licenses assigned thereto (Jisc, n.d.b).

Many challenges outlined in the following sections are specific to data sharing, and – since data sharing is the end goal of RDM – are challenges that also apply to RDM practices.

2.3.2 Challenges for Research Data Management

The following challenges highlight the logistical difficulties of RDM and data sharing, as well as factors that disincentivise researchers to share data:

- RDM is often affected by institutional, departmental, or disciplinary practices (Schöpfel & Prost, 2019: 98-99), making it difficult to develop standardised policies and training programs that can accommodate multiple subject areas.

- Complex decision-making regarding which data should be kept and managed – including how long and where to keep it – and which should be discarded (Beagrie, 2019:4).

- If data belongs to an external party, researchers will be restricted in terms of use and sharing. According to Borgman (2009:17), IP limitations are one of the main disincentives for researchers to share data.

- Sharing some kinds of data might also compromise individuals' identities or transgress privacy laws (Banks et al., 2018:4), and doing so could have ethical or legal implications. For example, researchers working with data under the remit of the European Union's General Data Protection Regulation (GDPR), which is "the toughest privacy and security law in the world" (Wolford, n.d.), may be restricted in their ability to access, re-use, and share certain data (The Open University, n.d.).

- Researcher anxiety regarding others finding fault with their data is another pressing factor (Nosek, 2012 in Banks et al., 2018:5). Although critical engagement is necessary for research integrity, it can result in conjectures that negatively affect academic reputations (Banks et al., 2018:5), and therefore disincentivise data sharing.

- Data might not be used or interpreted in the way researchers intended (Tenopir et al., 2011 in Banks et al., 2018:5; Borgman, 2012:1069) – a further contributor to disincentivised data sharing.

- A lack of reward or credit for researchers who share their data is another factor, where scholars' career development is rarely positively affected by sharing data (Borgman, 2012:1072; Borgman 2009:16). Given researchers' already pressing workloads, there is no motivation to invest the time and effort required to share data (Borgman, 2012:1072; Poole, 2013).

- A lack of funding and human resource capacities (Poole & Garwood, 2019:9), or capacities related to technical requirements of data collection, storage, and analysis (Poole & Garwood, 2019:5-6) are both factors as well.

- Finally, the question of which parties should be responsible for RDM and data sharing (Poole & Garwood, 2019:8) adds a layer of complexity. Whether, and to what degree, researchers, project managers, institutional libraries, funders, or publishers are involved is important to address.

2.3.3 Challenges for Research Data Management in the humanities

Due to the nature of humanities data, there are certain challenges specific to managing it:

- Humanities' RDM and data sharing infrastructure is less developed than that of the sciences, where humanities researchers are seldom encouraged or required to manage and share data compared to their counterparts in the sciences (Borgman, 2012:20).

- The sheer diversity of humanities data means it is challenging to standardise humanities RDM and data sharing practices, such as developing standardised metadata schema (Gómez, Méndez & Hernández-Pérez, 2016:547).

- Humanities data are often of a qualitative nature, while most existing RDM and data curation methods have been developed to accommodate quantitative data and research methods (Given & Willson, 2018:213; Munoz & Renear, 2011).

- Humanities data are often secondary in nature, where IP owned by external parties create significant obstructions in using and sharing data, especially since so much of humanities research is based on cultural records (Borgman, 2009:17).

- The humanities typically attract less funding than other disciplines, impacting the financial, technical, and human resources available for successful RDM implementation. Financial support is particularly important for the generation of persistent identifiers, providing metadata, creating and assigning licenses, data formatting, version history maintenance, and data peer review (Almas, 2017:1).

2.3.4 Challenges for Research Data Management in the Digital Humanities

Given the sheer size and interdisciplinary nature of Digital Humanities projects, engagement with the principles of RDM and data curation has been necessary from early on. Digital Humanities research relies on data ("What is Humanities…", n.d.) and thus presents an immediate need for data management, unlike the traditional humanities: "the digital humanities community […] already possesses sophisticated experience in preserving access to digital scholarship" (Munoz & Renear, 2011). The Digital Humanities have therefore already started pioneering development for the humanities in terms of RDM and data sharing (Flanders & Muñoz, n.d), but there is still much work to be done. RDM related challenges impeding this progress are discussed below.

- Despite the "interest in data curation in the digital humanities", Dressel (2017:5) notes that not much attention "has been paid to providing research data management instruction" to researchers in these fields.

- DMPs are often inflexible due to funder requirements, which, consequently, "may produce unanticipated outcomes" in relation to the DMP (Poole & Garwood, 2019:10), diminishing its purpose. This is particularly true for complex Digital Humanities projects with large teams and volumes of data. Additionally, funders do not tend to measure or ensure DMP compliance (Poole & Garwood, 2019:10).

- Digital Humanities project teams are often large and interdisciplinary, and difficulties faced in managing such teams (for example, translation work across linguistic borders, task delegation, and conflict resolution) relate to RDM as well (Poole & Garwood, 2019:9-10).

- Volumes of Digital Humanities data are often large, and need to be machine-readable (Wilms et al., 2019:27), meaning such projects often require specific technological resources that are not always available (Poole & Garwood, 2019:9-10).

- Digital Humanities projects also lack standardised practices regarding technical platforms and data structures (Borgman, 2008 in Borgman, 2009:15).

- A final RDM-related issue impeding data sharing progress in the Digital Humanities concerns inadequate funding. With such large teams, volumes of data, and technological requirements, researchers need adequate funding to run projects to completion where, as DARIAH (n.d.a) confirms, production and management of digital data in the humanities is "expensive, challenging and time-consuming".

2.3.5 Potential challenges for Research Data Management in Web Archive Studies

Potential challenges that researchers might encounter when establishing corpora relating more specifically to WAS projects are discussed below.

- Duplication of data in a web archive – usually caused by multiple hyperlinks pointing to the same web entity or web page – entails a time-consuming process of deciding which versions to include, and which to discard (Brügger & Laursen 2019:4).

- Lack of data in web archives is a further issue, and might occur when only certain parts of a web domain are included, when only small collections have been archived, or when there have been technical issues preventing all planned data from being archived (Brügger & Laursen 2019:4). In cases like these, a corpus might not accurately represent the web, and might combining smaller collections (Brügger & Laursen 2019:4).

- Since corpora – including the file types and formats – depend entirely on the needs of individual research project, they cannot follow a standard rubric, meaning there are ultimately no standard practice methods for archiving the web (Brügger, Laursen & Nielsen, 2019:124; Brügger & Laursen, 2019:5).

- The unsystematic nature of web archive material is troubling too, and requires meticulous, time-consuming data-cleaning processes carried out by various stakeholders, including curators and IT professionals (Brügger & Laursen, 2019:5).

- The diversity of digital data, and the consequently diverse ways in which it must be managed presents further complications (Brügger, Laursen & Nielsen, 2019:124).

- Finally, IP and third-party ownership limit the ways that researchers can create, share, use, and reproduce web data (Winters, 2019:80). Such limitations might even "distort" research, where legal restrictions ultimately affect researchers' "ability to explore and comprehend the histories" of national webs (Winters, 2019:80).

Given the various data management challenges within WAS, it is important for researchers in this field to keep clear records of their methods and processes (Brügger, Laursen & Nielsen, 2019:124), since "the ways in which things are collected, made accessible and documented have an impact on how they can later be used by researchers" (Schroeder & Brügger, 2017:9). If web archive data are to be shared, re-used, and/or validated, interactions with the data must be recorded appropriately. Although little extant literature discusses the importance of record-keeping in the field of WAS, the same does exist for humanities and qualitative research more generally, and which can be applied. Researchers such as Shenton (2004:72), Carcary (2009:11), and Bowen (2009:305) confirm that keeping meticulous records of research methods via an audit trail can be extremely valuable in contributing to the transparency and trustworthiness of a study when validity or reliability are hard to prove, especially since both concepts are often more relevant to quantitative studies.

2.5 Stakeholder support for Research Data Management and data sharing

Since the emergence of RDM imperatives, there has been incentivised uptake in terms of policy development, educational outreach, and system integration by various stakeholders in the research ecosystem. In an attempt to realise the benefits of sound data management, "funders, publishers, societies, and individual research groups have developed tools, resources, and policies to encourage investigators to make their data publicly available" (Piwowar, 2011:1). While stakeholders are

engaging with and supporting researchers' RDM and data sharing, there is still work to be done, and from various points of departure.

2.5.1 General support for Research Data Management and data sharing

There are a number of research organisations that provide RDM and data sharing support. The DCC and Jisc both provide training and educational resources. Guiding principles have also been created, such as the Transparency and Openness Promotion (TOP) Guidelines (Centre for Open Science [COS], n.d.c) and Findable, Accessible, Interoperable, Reusable (FAIR) principles (Wilkinson et al., 2016). A number of lifecycle models have emerged, providing structure for RDM and digital curation best practices. Tools for DMPs have been designed, such as DMPOnline (DCC, "DMPOnline", n.d.) and DMPTool (DMPTool, n.d.). The Research Data Alliance (RDA) further coordinates a number of working and interest groups in building "the social and technical infrastructure to enable open sharing and re-use of data" (RDA, "About RDA", n.d.). Various repositories, such as Dryad and Figshare, are also integrating with university systems and/or offering data curation services (Figshare, n.d).

In terms of RDM and data sharing support, the Consortium of European Social Science Data Archives (CESSDA) has released a Data Management Expert Guide advising best practices regarding qualitative and quantitative data, and provides assistance with various issues that arise from qualitative data types, such as privacy and copyright problems (CESSDA, 2020). Being European based, it also addresses the impact of GDPR on researchers.

Much of the institutional support for RDM and data sharing concerns the development of RDM policies, training, and researcher resources, and infrastructure installation. For example, the University of Edinburgh's library has a training course titled "MANTRA: Research Data Management Training" (MANTRA, n.d.) – an example of an early training course responding to RDM funder requirements. Institutional support for RDM can also be seen in projects like the Data Management Rollout at Oxford (DaMaRo), funded by Jisc, and tasked with creating a data management policy for researchers at Oxford University (DaMaRo, n.d.). DaMaRo has also released some case studies providing discipline-specific examples of RDM (Wilson, Patrick & Rumsey, 2013). Many libraries also manage their institutional repositories and promote advocacy and awareness regarding RDM.

In terms of support for the Digital Humanities specifically, organisations such as The Alliance of Digital Humanities Organizations (ADHO), which has a formal liaison with the RDA (ADHO, n.d.) and DARIAH, run a working group for RDM in the Digital Humanities (DARIAH, "Research Data Management Working Group", n.d.b). Various Digital Humanities centres have also emerged at institutions such as

the Centre for Digital Humanities at the University College London, HUMlab at the University of Umeå, and DIGHUMLAB in Denmark, and whose work partly involves data curation.

There is very little direct attention paid to RDM and data sharing for WAS specifically. However, one significant initiative involves a working group hosted by Web ARChive Studies Network (WARCnet) at Aarhus university, the aim of which is "to investigate the possibilities of sharing web archive-related data across borders" (Web ARChive Studies Network [WARCnet], 2020).

Although stakeholders have made some headway in terms of supporting researchers' RDM and data sharing practices, such support is often inadequate – particularly for researchers in the broader humanities and the Digital Humanities. Munoz and Renear (2011) argue that in order to promote sound RDM practices in the humanities, researchers must be supported through skills development, training, education, and through general institutional support with dedicated professional roles. Kvalheim and Kvamme (2014:7) recommend that formal funder policies and RDM processes be incorporated into the regular routine of humanities researchers. Beagrie (2019:14) calls on funders, repositories, institutions, and organisations such as Jisc, to collaboratively present workshops that "evolve disciplinary norms" on which data to keep and discard, especially for disciplines where such norms do not currently exist. Fear (2015:18) places importance on mentorship programs where researchers experienced in sound data sharing practices guide others in how they might develop such practices of their own. Simms et al. (2016) emphasise the importance of community collaboration in data management planning, since it promotes the expansion of outreach and training capacity. Buddenbohm et al. (2016:32) recommend that researcher training, the appointment of qualified staff, and the implementation of incentivisation systems are necessary in order to decrease researchers' reluctance to sharing data.

2.5.2 Publisher support for Research Data Management and data sharing

Publishers' involvement with RDM and data sharing is connected to the industry's general efforts to promote and encourage open science/open scholarship, where "open data sharing is a central element for making research more transparent, reproducible, and increasing its potential impact" (McKiernan et al., 2016 in Rousi & Laakso, 2020:132). According to Fear (2015:18), there is "an increasing emphasis on data sharing from publishers". It is no surprise then, that publishers' involvement in RDM and data sharing has, for the most part, concerned the development and implementation of journal data sharing policies. Citing Piwowar and Chapman (2008), Rousi and Laakso (2020:132) state that journal data sharing policies have been developing for over 20 years: the Association of Learned and Professional Society Publishers conducted a survey in 2008 showing that

around 47 per cent of publishers had a data policy for their journals (Herndon & O'Reilly, 2016:226). Some of these publishers include Oxford University Press, Cambridge University Press, Wiley, and Sage (Herndon and O'Reilly, 2016:230-234) – all of which can be considered early adopters of data sharing policies in the social sciences. For scientific disciplines, early adopters of data sharing policies include PLOS and BioMed Central (Hrynaszkiewicz, 2020), both of which currently require all authors to share their data (Association of College & Research Libraries, 2014).

Some of the bigger commercial publishers have released standardised policy frameworks requiring different levels of adherence, including Taylor & Francis (Taylor & Francis, n.d.c), Springer Nature (Springer Nature, n.d), Wiley (Wiley, n.d.b), and Elsevier (Elsevier, n.d). The more basic policies tend toward encouraging data sharing, while stricter policies might require that data are shared according to FAIR principles. For example, Springer Nature's (n.d) 'Type 1' policy encourages the sharing of data, while its 'Type 4' policy requires it. A number of publishers are members of the RDA, which has an interest group specifically focusing on data policy standardisation and implementation (Research Data Alliance [RDA], n.d.b).

Since most research findings are published in scholarly journals, journal data sharing policies have the potential to influence how researchers make their data available (Rousi & Laakso, 2020:132). Jones, Grant, and Hrynaszkiewicz (2019:3) posit that such policies allow publishers to "inform, support and encourage authors to share the data underpinning their research". Indeed, publishers' data sharing policies have been identified as key to encouraging and ensuring that data are shared by researchers (Sturges et al., 2015:2447). Although there are clear benefits to rolling out such policies, doing so also presents challenges, particularly in terms of researcher compliance.

The Primary Research Group (2018:21) conducted a survey showing how researchers are affected by data management requirements. Only around 47 per cent of researchers stated that journal publishers' data requirements impacted their work (The Primary Research Group, 2018:21). Christian et al. (2020:1) found in their study – which examined and compared publisher data sharing to editor and author understandings of the same policies – that although the number of journals with data sharing policies have increased, researchers do not always share their data, and when they do, the data are not of adequate quality for verification or re-use. Naughton and Kernohan's (2016:84) study assessing the possibility of a data policy registry to assist researchers with policy compliance found that if such a registry were built, the entire "research data policy ecosystem" would be "in critical need of standardisation". Vines and Albert (2020) noted that researcher submissions to journals without a data sharing policy seemed to increase slightly over time, while journals with strict data sharing

policies saw a decline in submissions, indicating authors' reluctance either to share data, or to doing so correctly.

In recent years, publishers have also become involved in RDM and data sharing outside of policy development. First, publishers have begun to offer – and sometimes require – a DAS, which "tells the reader where the data associated with a paper is available, and under what conditions the data can be accessed" (Taylor & Francis, n.d.a). DASs are often referred to in data policies. For instance, PLOS has required DASs for every published article since 2014 (Federer, 2018).

Second, many publishers have adopted Open Science Badges, an initiative by COS (n.d.a). The badges visually acknowledge data sharing undertaken by a researcher in a published article, and aside from incentivising researchers through acknowledgement, intentionally signal to other researchers that data has been made available with a persistent identifier (COS, n.d.a). No studies have been conducted yet regarding the impact of said badges on researchers' data sharing behaviours.

Third, publishers offer some general guidance to researchers in sharing their data. For example, Taylor & Francis has a webpage titled "Understanding Our Data Sharing Policies" (Taylor & Francis, n.d.c) that includes guidance on which repository to choose, how to cite data, how to fill out a DAS (including a DAS template), and a FAQs section, amongst other information. Springer Nature has a webpage titled "Research Data Policy Types" (Springer Nature, n.d.), which offers similar guidance along with a research data helpdesk that researchers can contact. Such guidance is, however, quite general and does not address the intricacies of Digital Humanities data or offer the kind of support that researchers might need in these fields.

Fourth, publishers are beginning to involve themselves in RDM training. Elsevier has partnered with a group of university libraries to launch The Research Data Management Librarian Academy (RDMLA): an online professional development program for librarians and other information professionals teaching the skills needed for effective collaboration with researchers about their data management (Research Data Management Librarian Academy [RDMLA], 2019). Although this training is meant for librarians specifically, it ultimately affects the training and skills of researchers by extension. Importantly, RDMLA "is a unique partnership between librarians, LIS educators, and a publisher" (Shipman and Tang, 2019:243) – that is, there are no other similar collaborations that include publisher perspectives. Given that RDMLA was formed in 2019, there are not yet any studies on the efficacy of the program.

Fifth, publishers are collaborating with external organisations as well as each other. As mentioned earlier, some publishers have collaborated with the RDA in a data policy standardisation interest

group, and a research data policy framework has since been released (Hrynaszkiewicz, 2020). As part of general efforts to further the objectives of open science/open research, publishers are also beginning to install departments and invest in human resource capacity dedicated to open research initiatives. Publishers with such departments include F1000, Springer Nature, PLOS, and Taylor & Francis, among others.

Some of the above publishers such as BioMed Central (BioMed Central, n.d), Wiley (Wiley, n.d.a), F1000 (F1000, n.d), and Taylor & Francis (Taylor & Francis, n.d.b) also accommodate the publication of registered reports. According to the COS, registered reports "[emphasize] the importance of the research question and the quality of methodology by conducting peer review prior to data collection" (COS, n.d.b). The facilitation of registered reports therefore allows publishers to indirectly support researchers' data planning and research methods.

One of the difficulties that publishers grapple with is how to support researchers, especially in terms of data sharing in across the broader humanities (and by extension the Digital Humanities). Jones, Grant, and Hrynaszkiewicz (2019:4) note that, although humanities researchers welcome the idea of open research, sharing, and collaboration, "the relationship that humanists have with their underlying research resources differs from that in other disciplines, and this needs to be reflected in data management and sharing policies". In an effort to address the lack of differentiation between humanities data from other disciplines, a cross-publisher working group was established in 2020 – an output of the International Association of Scientific, Technical, and Medical Publishers' (STM) initiative 'Research Data Year' – focusing on supporting humanities researchers in managing and sharing their data. This working group has brought together professionals from multiple publishers interested in data management and sharing in the humanities and aims to develop resources including case studies. Further details of the working group are not yet available since they were only established in the last quarter of 2020.

As illustrated in the literature discussed above, much of the work publishers have been doing regarding RDM and data sharing support concerns data sharing policy development and implementation. While there is information and advice provided by some publishers regarding the sharing of data, there is an evident lack of guidance offered for humanities researchers specifically.

2.6 Chapter conclusion

In concluding this literature review and moving on to discussion of methods, the following question posed by Borgman (2009:21) should be kept in mind: "Why is no one following digital humanities scholars around to understand their practices, in the way that scientists have been studied for the last

several decades?" While it is clear that some progress has been made in this area since then, there is more work to be done, which this study aims to contribute to. As Borgman (2009:21) explains, such investigation into research practices "has informed the design of scholarly infrastructure" for the sciences, and so it should be for the humanities too.

Chapter 3: Research methodology

In this chapter, I consult a number of methodological texts appropriate to the humanities and social sciences in support of my decision to conduct the research as an exploratory case study.

3.1 Research paradigm and approach

Broadly speaking, this study is constructivist or interpretivist[4] in nature, rather than positivist or post-positivist. While a positivist approach might be appropriate for testing an extant theory, a constructivist approach emphasises subjectivity in terms of both data collection and analysis. That is, the researcher is not considered to be 'outside' of the research (as with positivism) – their reality plays a part; thus the researcher 'interprets' what they find, just as their subjects present information according to their own experiences and interpretation thereof (Bhattacherjee, 2012:103).

The current study uses both inductive and deductive approaches. As Bhattacherjee (2012:103) explains, an interpretive approach to research often involves developing or inducing conclusions about phenomena through the process of data observation. Given that there has been little peer-reviewed literature published on the topic of RDM in the field of WAS, and consequently no standardised process that researchers might follow, this study is exploratory in nature. As Babbie (2016:90) notes, social research is often "conducted to explore a topic", where an inductive approach can be particularly useful in studies investigating new subjects. However, since the current study refers to a framework as well as relevant literature which the collected data will either confirm or refute, it also deduces conclusions about the research. Both Swain (2018) and Fereday and Muir-Cochrane (2006) posit that hybrid methodologies offer a useful approach to thematic or content analyses. Specifically, Swain (2018) outlines a hybrid approach as using a "top-down, deductive, theoretical process" to produce a set of *a priori* codes based on the proposed framework and research questions, and a "bottom-up, inductive, data-driven process" to produce a set of a *posteriori* codes "derived from an examination of data generated".

The study uses a convergent mixed methods approach, 'convergent' design being used when both qualitative and quantitative data collection methods are equally prioritised, are implemented concurrently, and the results combined during data analysis and interpretation (Creswell and Plano Clark, 2011:70). Based on definitions by Meyer (2001) and Harrison et al. (2017), quantitative data are numerical, and qualitative data are not; both are useful, but both have limitations. On one hand, quantitative data can lack the in-depth detail that qualitative data might provide, and can thus lack

[4] The interpretive paradigm is also referred to as the constructivism paradigm (Kivunja & Kuyini, 2017:33).

"richness of meaning" (Babbie, 2016:26). On the other hand, qualitative data can produce ambiguity in a study given that it is "purely verbal" rather than statistical, and relies on subjective interpretation (Babbie, 2016:26). A mixed methods approach employing both qualitative and quantitative data ultimately "makes for stronger research" and allows for diversified data (Babbie, 2016:27). By diversifying a dataset, the study in question "may lead to unique insights" that would not have been possible otherwise (Bhattacherjee, 2012).

3.2 Research design

The chosen research design concerns a case study, which, commonly, is "most suitable for a comprehensive, holistic, and in-depth investigation of a complex issue […] in context" (Harrison et al., 2017). As Stake notes, case studies "facilitate the understanding of something" and are "instrumental in providing insight on an issue" (Stake, 2006 in Harrison et al., 2017). While the current study investigates data management and sharing, and how publishers support these practices, it does so through a specific case lens of WAS. The case will help gather insight into the more overarching field of Digital Humanities.

3.2.1 Strengths and weaknesses of case study design

As Meyer explains, "there are virtually no specific requirements guiding case research. This is both the strength and the weakness of the approach" (2001:329). Case research allows for flexibility and the opportunity to choose data collection methods most appropriate to the study at hand (Meyer, 2001:329). Such flexibility, however, can result in case study research being constructed according to a "rather loose design" which lacks structure and robustness (Meyer, 2001:330). To counteract this, it is necessary to keep meticulous records and ensure the reliability and validity of case study research (Meyer, 2001:329 and Harrison et al., 2017).

3.3 Research methods and instruments

There were two main methods used to collect research data for the case study: semi-structured interviews and an online questionnaire.

3.3.1 Interviews

Interviews were selected as a data collection tool since case study research supports in-depth and detailed understandings of phenomena through subjective experiences and interpretations thereof. Babbie (2016:311) confirms that interviews allow researchers to gather qualitative and subjective detail that would not be possible otherwise. In this case, the interviews allowed for a detailed

understanding of the activities publishers are currently engaged with regarding facilitation of RDM and data sharing, and an equally detailed understanding of WAS researchers' RDM practices and the kinds of support they seek from publishers.

Semi-structured interviews, more specifically, allow for a degree of flexibility since they consist of a set of open-ended questions, but allow for follow-up or probing questions that arise from interviewees' responses (Morse, 2012). The interviews were conducted and recorded remotely using Microsoft Teams.

3.3.2 Questionnaire

Questionnaires can be used in various ways and prove to be suitable tools for many different research purposes, including those that are exploratory in nature (Babbie, 2016:254). In answering the current research questions, it was important to gather responses and viewpoints from multiple researchers within WAS and the Digital Humanities using a questionnaire, rather than focusing on just a few select perspectives via the interview instrument. The intended population were geographically scattered and mostly unknown to the current researcher, making an online questionnaire the most efficient way of gathering data from them.

In accordance with guidance set out by Babbie (2016:249), the questionnaire was constructed using a mix of appropriate question types. Each question item was short and clear, negative items or biased terms were avoided, and all were ordered and formatted in a way that emphasised clarity. The questionnaire predominantly collected quantitative data through closed-ended questions, but included some qualitative items when respondents were given the option to answer open-endedly.

The questionnaire was created and formatted using SurveyMonkey. It was distributed via the mailing lists or newsletters of five groups and organisations, and via the social media platforms of a further two organisations, all of which have focused interests in either WAS or the broader Digital Humanities.

3.4 Population and sampling

3.4.1 Population

As Babbie (2016:117) explains, the "population for a study is that group […] about whom we want to draw conclusions". The current study involves two different populations. The population for the questionnaire and one of the interviews concerned researchers situated in the field of WAS. The population for the remaining two interviews was that of commercial publishers who engage with data management or data sharing.

3.4.2 Population size

As there is no dedicated group of WAS researchers in South Africa, it was necessary to extend the gaze internationally. There being no definitive database of WAS researchers, the population size for this group has been calculated according to the membership, subscription, and follower numbers of the following organisations: The Digital Humanities Summer Institute (DHSI), the International Internet Preservation Consortium (IIPC), the Web ARChive Studies Network (WARCnet), DARIAH, Archives Unleashed, the Digital Preservation Coalition (DPC), and the Engaging with Web Archives (EWA) Twitter page. Each of these organisations are either WAS, digital preservation, or Digital Humanities focused, and the total number of members and subscribers to these organisations is 8168.[5] A limitation with this calculation is that the membership and subscriber bases of the organisations were not all researchers in WAS. The calculation, therefore, can be considered accurate in terms of the number of questionnaire recipients, but perhaps not so in terms of the number of actual WAS researchers.

To obtain a more accurate idea of the number of potential WAS researchers within the population of 8168, the number of authors who published in a journal in the field of WAS between 2017 and 2019 were considered, as well as the number of authors from five broad-scope, multi-authored volumes on the subject of web history published between 2010 and 2019. To account for the researchers not represented in these publications, 30 per cent was added to the total, producing the outcome of 229 researchers. While this number is only an estimate, they illustrated that the population herein was sizable enough to conduct sound research.

Similar to WAS researchers, the population of publishers actively engaged with data management and sharing is scattered, with no definitive databases to search. However, many of these publishers are signatories to the TOP Guidelines – an initiative developed by COS as a resource for the development and assessment of data sharing policies. COS has published a list of all organisations that are signatories on their website (COS, n.d.c), indicating 33 commercial and society publishers.[6]

[5] The DHSI has 5444 subscribers to their mailing list. The IIPC has 660 subscribers to their mailing list. WARCnet has 63 subscribers to their mailing list. DARIAH has 667 subscribers who receive a monthly newsletter. Archives Unleashed has 339 members on their Slack channel. The DPC has 578 subscribers to their mailing list. The EWA Twitter page has 417 followers. The subscriber numbers were provided via email by each organisation between September and October 2020 whilst seeking consent to distribute the questionnaire via the relevant mailing lists, newsletters, and social media channels.
[6] The number of publisher signatories of the TOP Guidelines was accurate as of 15 September 2020.

3.4.3 Sampling method

It is usually impossible to study all the units in a given population, and it becomes necessary to select a smaller sample as a representative of the wider population (Babbie, 2016:117). The questionnaire and the interviews herein both used non-probability sampling techniques given that "some units" of the respective populations had "zero chance of selection" (Bhattacherjee, 2012:69). In other words, not every person in the populations responded to the questions being asked.

For Web Archive Studies researchers to whom I wanted to distribute the questionnaire, self-selection sampling was used, where sampling units choose either to participate or not of their own accord – a technique best employed for "difficult-to-locate" populations (Sterba & Foster, 2008). Since the study at hand sought to sample scattered researchers in an emerging field (not all were included in the same database or subscribed to the same organisation), the self-selection sampling technique was necessary. Participants self-selected by assessing whether they were eligible to complete the questionnaire after receiving the link via email (disseminated through a mailing list, newsletter, or a social media channel).

Purposive sampling was used to select one researcher, and publishers, to interview. This is a technique "in which the units to be observed are selected on the basis of the researcher's judgment about which ones will be the most useful or representative" (Babbie, 2016:188). The interviews were conducted with participants chosen for their expertise in their respective fields.

3.4.4 Sample size

Regarding the publisher sample, interviews with two parties were conducted: a specialist in open scholarship responsible for the publishers' data management and sharing activities; and a portfolio specialist in the fields of history, media, and the arts. Considering the scattered nature of the publisher population, and since the study is exploratory in nature, two publisher participants were identified whom the researcher knew would willingly participate due to their engagements with data and data sharing. As Vasileiou et al. (2018) note, "[s]amples in qualitative research tend to be small in order to support the depth of case-oriented analysis that is fundamental to this mode of inquiry".

For the WAS researcher sample, one interview was conducted with a professor in media studies who has an academic focus in web studies, and is the editor of a journal in the field, and who was identified due to their WAS activity and research. The questionnaire sent to WAS researchers returned 31 self-selected respondents.

Although the sample sizes are small, this is often the case for exploratory research where a new topic or field is being investigated, as is the case herein.

3.5 Study validity and reliability

Reliability and validity play an important part in ensuring the accuracy of research (Creswell, 2018:199). According to Babbie (2016:146), reliability "is a matter of whether a particular technique, applied repeatedly to the same object, yields the same result each time". Reliability, as a measurement, is not appropriate for the current study's qualitative interview data because of their semi-structured nature. In the questionnaire, however, reliability was pursued through carefully constructed questions. Following the advice of Bhattacherjee (2012:56): the questions were generated in a way that avoided dependency on subjective answers; respondents were only asked questions they would likely to be able to answer; the questions did not include any ambiguous items; and the questions used only clear and concise language to avoid misinterpretation on the part of respondents.

Validity is a technique that measures that which it intends to measure (Babbie, 2016:151; Creswell, 2018:199), and is relevant to both the quantitative and qualitative aspects of the current study. Linked to the concept of validity is that of trustworthiness, which pertains to the study's qualitative aspect. For interpretive studies, Angen (2000:392) suggests that "the notion of validity as truth or certainty must be abandoned" and understood rather as "an evaluation of trustworthiness".

Validity was maintained in the following ways:

- All questions (within interviews and the questionnaire) were structured according to a trusted data lifecycle framework – the JRDL – ensuring that all lifecycle stages and processes typical to a data lifecycle were addressed by the questions.

- A pilot study of the questionnaire was conducted: the questionnaire draft was sent to five of the researcher's peers to test for clarity.

- The final interview and questionnaire templates are openly available to readers in Appendices A, B, C, and D of this study.

- Once the interviews had been conducted, the interview transcripts and recordings were sent to each interviewee for a member check, and each interviewee confirmed the accuracy of their responses.

- The study is triangulated in that it generates data from more than one source (researchers and publishers) on the same topic, and uses a mixed research method (qualitative and quantitative data) with different research instruments (interviews and a questionnaire).

- An audit trail of all interview and questionnaire responses and recordings was maintained, along with any changes requested by peers or participants in a research diary.

Trustworthiness can be maintained by keeping a written account of the research, thus recording the process through which conclusions are drawn, and maintaining transparency regarding this process throughout (Angen, 2000:390). This approach "should allow the researcher to face criticisms of subjectivity—of this being just their opinion or even just the opinion of their participants—with evidence of what has been brought to bear on the interpretation" (Angen, 2000:390).

3.6 Ethical considerations

This study involved gathering information directly from human subjects. As such, ethical clearance was required from the Ethics Committee of the Faculty of Humanities at the University of Cape Town.

The researchers had no previous relationships with any of the interviewees – first contact was made with selected interviewees via email where they were informed of the current study and their willingness to participate was enquired after. Each interviewee stated their willingness to be interviewed. Formal written consent was also sought from each interviewee, regarding the recording of the interviews, and in terms of referencing details of their employment and/or other official positions. Consent was sought via signed consent forms (see Appendix E), which were sent along with a list of the interview questions to each interviewee two weeks before the interviews were scheduled to take place, allowing ample time for the forms to be signed and returned, and for participants to prepare for the interviews. The interviews could only be conducted after receiving the signed consent forms and confirmation regarding participants' willingness to answer the proposed questions. For the purposes of confidentiality, interview participants are not named herein, and their recorded interviews not be made publicly accessible.

The questionnaire did not collect identifying details, maintaining respondents' anonymity. Babbie (2016:65) resonates here: questionnaire respondents' anonymity will "increase the likelihood and accuracy of responses". Although the processed questionnaire data is referred to in the study results, the data itself will not be made publicly accessible. These assurances were made to each respondent in the questionnaire preface, and although there were some demographic questions, these were only

to ascertain whether the data could be used or discarded. No information that could potentially reveal the identities of the questionnaire participants is referred to in the current study.

No incentives or direct benefits were advertised to either the interview subjects or questionnaire respondents to encourage participation. An indirect benefit for the commercial publisher interviewees is that the proposed recommendations following the outcome of the current research might positively influence their engagements with researchers. The indirect benefit for the other interviewee and the questionnaire respondents (and all researchers in WAS) is that commercial publishers may increase or improve the data management and sharing support that they provide researchers in the Digital Humanities.

The current researcher is employed by Taylor & Francis, who have partially funded this study. No directive was received from Taylor & Francis regarding the topic and study process, though the company may benefit from the study findings in the same way that other publishers might.

3.7 Data collection

A list of questions was drafted for the questionnaire and then transposed to SurveyMonkey and made available online. After receiving permissions from the organisations mentioned in section 3.4.2, the questionnaire was distributed throughout their mailing lists, newsletters, and Slack channels. Given that the initial distribution phase yielded few responses, the questionnaire was distributed a second time. Responses to the questionnaire appeared in the SurveyMonkey dashboard as and when they were submitted. When the questionnaire closed, responses were exported into a Microsoft Excel spreadsheet for analysis.

In preparation for the interviews, lists of interview questions were outlined. Each interview schedule differed slightly, and have been made available for reference and re-use in Appendices A, B, and C. Each interview was conducted using Microsoft Teams – a communications application that allows for video conference calls and audio recordings. This was necessary since the interviews were conducted remotely – the three interviewees are all based in Western Europe while the current researcher is based in South Africa. The interview recordings were saved as MP3 audio files and transcribed into a text document.

The data collection stage took place over a four-week period in November 2020, allowing time for interviews to be scheduled, recorded, and transcribed, and for the questionnaire to be distributed twice.

3.8 Data analysis

The analysis technique used to assess the qualitative interview responses was that of content analysis, involving a process of textual coding to ascertain common themes (Bhattacherjee, 2012:115). The transcribed interview data was coded according to pre-decided thematic categories based on the structure of the JRDL model. The interview responses revealed emergent themes that were outside the scope of the pre-determined categories. In order to maintain transparency and present the data accurately, direct quotations from the interviews are referenced as support for said categories during discussion in Chapter 4.

The analysis technique used to interpret the quantitative questionnaire data was inferential statistics, which are "the statistical measures used for making inferences from findings based on sample observations to a larger population" (Babbie, 2016:460). Further tools used included univariate analysis, which is the "analysis of a single variable, for purposes of description" (Babbie, 2016:417), allowing the study to refer to measures, such as the percentages of researchers who partake in each stage of RDM.

3.9 Limitations

There were a number of limitations to the methodology. First, I was not able to conduct in-person interviews as the interview subjects were all based in Europe while I am based in South Africa. Second, the populations of both WAS researchers and commercial publishers currently engaging with data management and data sharing are both small, and the sample sizes also small due to the field being very new. Third, there is no definitive database that includes all WAS researchers, making definitive population numbers impossible to ascertain. How these limitations affected the current study will be discussed in Chapter 5.

3.10 Chapter conclusion

This chapter outlined the research paradigm as interpretivist, with both an inductive and deductive approach to the current study. It discussed qualitative and quantitative approaches and explained why a convergent mixed method was most suitable. The chapter explained why a case study – exploratory in nature – was chosen as an appropriate research design. The chapter then described the research instruments used, including interviews and a questionnaire. The chapter also explained how the populations were calculated, and how the sampling was carried out. The chapter outlined how the data analysis was done, as well as the research ethics, limitations, and the reliability and validity of the case study.

Chapter 4: Data analysis

This chapter presents data gathered from three interviews and a questionnaire, where responses include perspectives from both publishers and WAS researchers. The questions posed were intended to address the research objectives, namely: to provide details on the nature of data and the current RDM and data sharing practices of researchers in the field of WAS, and to discover the ways in which publishers are currently engaging with and supporting WAS researchers with their RDM and data sharing.

Questions posed to WAS researchers aimed to elicit information regarding the nature of the data they use, their current RDM and data sharing methods and practices, and to identify extant gaps in their RDM practices when compared to the stages in the JRDL. The questions posed also sought to identify what key challenges said researchers experience in managing and sharing their data. Considering that one of the objectives of the study aims to recommend ways that publishers might better support WAS researchers regarding data management and sharing, researchers were also questioned regarding support they might have received from publishers in the past, and support they feel would be useful in the future.

Questions posed to publishers intended to elicit information regarding data management and sharing support that publishers currently offer WAS researchers. The questions further sought to identify the roles that publishers see themselves playing in supporting researchers' data management and sharing, and the kinds of support that publishers might not offer currently but hope to in the future.

The in-depth interviews yielded a richer data-set than the questionnaire, due to the small sample of questionnaire respondents, and since the interviews included responses from both publishers and a researcher. Thus, the interview data are presented first, followed by the questionnaire results.

4.1 Interview data

Two of the interviews were conducted with employees at a commercial academic publisher, hereafter referred to as 'R1', and 'R2', and one interview was conducted with a WAS researcher, hereafter referred to as 'R3'.

The interview data forms the qualitative part of the data analysis. Responses from all three interviews are discussed together, and are organised thematically according to the following sub-categories: WAS as an emerging field; the nature of WAS data; RDM and data sharing practices of WAS researchers

presented and sub-categorised according to the JRDL research stages; publisher support and responsibility; and stakeholder collaboration.

As mentioned earlier, while this case study focuses on WAS, certain aspects about the nature of data, data management and sharing practices, will be detailed for the broader Digital Humanities and humanities more generally, but only as is relevant to the field of WAS. This is important to mention, since much of the interview data refers to the humanities more generally.

4.1.1 Web Archive Studies as an emerging field

R2 defined WAS as a "constantly evolving" area of research "that not only uses web archives as a primary source of evidence and inquiry, but focuses on some of those broader societal and cultural challenges and opportunities that might be associated with the preservation of and access to materials that have their origins on the web".

R3 defined WAS as research focusing on the web in its archived form. R3 noted that, although the web has existed for decades, it has only been viewed since the early 2000s as a research object in its own right, meaning the field is therefore still developing.

4.1.2 Data in the humanities and Web Archive Studies

R2 noted that many humanities researchers would not consider their research sources to be data, and that 'data' might not be an entirely relevant or useful term for them since many data sources are already codified in a given study's literature review or reference list. R2 pointed out that perhaps a different term is needed for the humanities, but that this term would likely also be limiting and problematic. R1 held a similar opinion: "some researchers may consider what they are collecting as 'data', other researchers will not". R1 questioned whether the humanities should employ the term 'data' at all, or whether a different term should be used; an alternative would be to refer to the data as its type: "a photograph, a recording, some sheet music, an annotation, but […] that partly contributes to some of the hesitation around data sharing because people don't recognise their annotation as data" (R1). R1 also noted that their employer allows authors to publish supplementary material alongside their published journal articles, but that "some of the lines between what is supplementary material and what is data can be quite blurry", particularly for humanities (R1).

R3 noted that for many research communities, data are understood as objects that can be quantified numerically. In the context of WAS, a web archive could be understood either as a dataset or as an

historical source (R3). R3 further noted that "the first [data sharing] obstacle" to overcome in said field "is to have people consider what they study *as* data".

4.1.3 Data management planning

The submission of DMPs by authors is not currently a common requirement when publishing in a journal (R1). However, there is a white-label multi-level data sharing policy being developed as an RDA output that publishers can adopt, that "ask[s] authors to share details" of their DMP (R1).

R2 noted a lack of funding for humanities subjects, and a consequent lack of funder mandates for DMPs and sharing data (that researchers in the sciences would have). R2 also noted that DMPs might not be as valuable to humanities research as making its methodology transparent, particularly regarding data sharing and research publication.

From a researcher perspective, DMPs do offer some value, but are difficult to continuously update due to unplanned changes in the data analysis process, for example, when conducting WAS research (R3). R3 confirmed that they do not create DMPs due to their research not being funded by an organisation that mandates DMPs.

4.1.4 Data collection and processing

4.1.4.1 Data types, formats, and sizes

R1 noted that, while not available currently, they hope that in the near future their employer will provide authors with more data sharing guidance based on humanities data types, regardless of their variability. For WAS specifically, R2 stated that data might include textual lists, website links, or screenshot images.

The data used for R3's research is all extracted from a pre-existing web archive, where R3 noted two main data formats: CSV files which can be processed using R[7]; and WARC files, which can be processed using other analytical tools. R3 stated that volumes of data can range from megabytes to terabytes depending on the project.

[7] R is "a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS" (R Foundation, n.d).

4.1.4.2 Time taken to collect data

R1 noted that humanities researchers might take years or even decades to generate a data-set that then informs multiple publications. For R3, it took about four or five months to extract the data needed for one of the more complex WAS projects they worked on, due to legal and technical issues.

4.1.4.3 Web archive data extraction

R3 noted one of the main challenges of extracting data from web archives concerns legal restrictions around access, permissions, and security. Further details of such restrictions are presented in section 4.1.5.4. When R3 *is* permitted to access and extract data from a web archive, there can be additional complications. R3 referenced a project where the data extraction was complicated because the manner in which web content is archived changes over time, meaning different approaches to data extraction are required, often within the same research project, depending on the historical timeframe (R3). Web archives are both complex and messy, and in order to extract the correct data, one needs to have a solid grasp of the archive's set-up and how to make good use of it (R3). R3 noted that in order to process the data for the project referenced earlier, a high-performance computer was required. They also noted that when data are extracted in CSV format, data cleaning is often a necessity (R3). R3 admitted that if the research team had realised from the start how difficult it was going to be to collect the archived web data needed, they might not have initiated the project at all.

4.1.4.4 Documenting methodology and metadata

R2 placed greater importance for humanities researchers to record their research methodologies than on sharing their data. R3 noted the crucial importance of meticulously documenting research methodologies in a logbook in terms of what data are extracted, how the extraction is performed, and how the data are used. R3 usually attaches metadata to the dataset that describes how the data was extracted, and how to use the dataset, further using file naming conventions including numbers, dates, and descriptive keywords that reference descriptions of the research stages for each data entry. If a mistake is made, it would be difficult to make adjustments to the data processing approach if there was no record of what was done in the past (R3). For WAS, much of the data work is iterative, which requires testing methods and assessing their efficacy (R3). R3 mentioned that documenting methods is particularly important for a new discipline where researchers are "making the road while driving on it".

4.1.4.5 Researcher team and collaboration

R3 confirmed that in the field of WAS, research projects are conducted both solo and by teams containing differing skill sets and competencies. One previous project involved multiple researchers generating and processing the data, but with each using the same data in different ways (R3). It was noted that working within a research team was often complicated due to differing skillsets, approaches, and expectations; for instance, researchers needed to learn how IT developers work, who often use iterative and agile processes (R3).

4.1.5 Data sharing

4.1.5.1 Data sharing infrastructure

R1 spoke about how the infrastructure in the sciences that facilitates data sharing is far more advanced than in the humanities, stating that "we're still really in the early stages of talking about what data sharing means for the humanities", so the infrastructure being relatively undeveloped is "not surprising". When asked if researchers in the humanities need to be supported differently to those in the sciences regarding data management and sharing, R1 answered "yes", explaining that there is a need to provide "specific guidance for humanities researchers on how to share qualitative data". There are specific challenges relating to humanities data collection not currently addressed by many of the resources made available by publishers, "because they've been more geared towards the life sciences" (R1). R2 also confirmed that since most shared data "isn't coming from […] humanities communities," there is less infrastructure to accommodate and support it.

4.1.5.2 The importance of data sharing

R2 noted that data sharing can be important for the field of WAS, but that for some humanities researchers it may be unnecessary since data management and sharing are already addressed by literature reviews and the maintenance of citations and reference lists. Applying data sharing practices that are best suited to other subject areas is not always relevant to the humanities (R2). R3 held a similar opinion: that although maintaining meticulous record of how data were collected and used is crucial, "in many cases, you don't actually need to share the data" because there is already an audit trail.

Similarly, R2 noted that instead of making humanities data itself available and accessible to others, it might be more important to emphasise research methodology transparency by publishing or linking to it in journal articles or books. R2 proposed the idea of using multimedia to showcase this, rather

than creating a blow-by-blow textual audit trail. They also noted the value of making research instruments such as survey questionnaires available as part of humanities research (R2).

4.1.5.3 Data sharing policies

The publisher employing R1 offers a suite of five journal data sharing policies ranging from encouraging to requiring researchers to share their data and fill out a DAS linking to the data. R1 stated that the majority of the journals published by the company began by adopting the most basic data sharing policy offered, but that this "hasn't led to very many authors sharing their data". There are active efforts being made to roll out stricter policies where possible, but the ability to do so is dependent on subject area (R1). Although the suite of policies offered was intended to be "subject neutral" given that different subject areas have different data needs, R1 confirmed that adjustments are being made to the policies based on feedback from researchers based largely in the sciences. R1 clarified that although most of the guidance provided to authors is based in the sciences, "that is not to say we're not interested and actively working on more specific guidance for humanities areas, it's just not available at this time".

Regarding the humanities specifically, R2 noted that "there is no point" in publishers "putting a policy in place that's not going to be acted on," especially without providing adequate support to researchers on how to adhere to such policies. R1 further highlighted the difficulty of complying with strict data sharing policies when a journal simultaneously requires double-blind peer-review: a published dataset would usually include the data creator's name, which would compromise any peer-review process requiring author anonymity.

For example, an internet studies journal – published by R2's employer – serves a community that makes use of web archive data and employs a very basic data sharing policy, encouraging but not actively seeking data sharing (R2).

4.1.5.4 Copyright, privacy, access, and security

From a publisher's perspective, legal restrictions to data sharing have been flagged as a major concern for the humanities in general, and for WAS. R2 stated that "there are some real issues […] around privacy and access", which are linked to issues around copyright. For R2, "if there are restrictions on what can be shared out of a particular web archive", this "presents a barrier for authors". R1 confirmed that some humanities data types have far more complicated ownership issues than some life science data types, and that humanities researchers wanting to share their data will "need very specific guidance and advice" on how to navigate the legal implications.

R3 confirmed from experience that data sharing is incredibly difficult in WAS due to legal restrictions, particularly regarding privacy, copyright law, and GDPR regulations, given their research is based on data from a European web archive. R3 stated that it took over a year for a contract to be drawn up that allowed the extraction of data from the web archive, which could then be stored on their university's server and used only under strict conditions. Although the extracted data that R3 uses cannot be shared openly, the contract template is available for sharing, re-using, and adapting by other web archives and researchers. In addition, R3 highlighted that different countries have different legal restrictions that may further complicate access and permissions, noting that the inability to share data may come down to just one URL link that includes a copyrighted brand name. They stated that it might be difficult to determine who holds the rights in the first place (R3).

R3 stated that, currently, the only way that data sharing would be possible in the above circumstances is if the data extraction was placed in a repository but protected from public view. If another researcher wanted to view or use the data, they would need to sign a separate contract with the web archive custodians (R3). However, such a repository does not yet exist. Additionally, the current contract that allows for data extraction also requires the data to be deleted after five years, which limits its re-use potential (R3). R3 confirmed that legal restrictions have thus perhaps been the main obstacle in their research.

4.1.5.5 Data licensing

R1's employer provides information and support to authors regarding data licensing, which becomes more important the stricter data sharing policies are. R1 mentioned that in future, they hope for more clarity around licensing humanities data specifically, including WAS. R1 further noted that supplementary material can also elicit data licensing issues since the same license that applies to a journal article is applied to its supplementary material.

4.1.5.6 Disincentives for sharing data

Given that humanities researchers often take much longer to generate their data, and might publish multiple pieces of research from the same dataset, R1 noted that researchers might be disincentivised to share their data due to the risk that someone else could use the key points in the data before the researcher has had a chance to do so themselves. R2 mentioned that humanities researchers might also be disincentivised to share their data or research instruments due to fear that they are used out of context. R3 noted that the legal difficulties with data extracting and sharing can be a major disincentive to researchers due to the complex, and sometimes costly nature of securing permissions.

4.1.6 Storing and archiving

There is currently no repository for the long-term preservation of web archive data. Although it would be extremely helpful to the community, there are complications regarding who would curate and fund such a repository (R3). Without the development of a repository to preserve extracted data, the time and money spent on creating data sets will essentially be lost after the period of time for which the data are legally contracted for use (R3). In terms of more informal data sharing, this would usually only be possible between the research team members who have contracted rights to use the data (R3).

Instead of using a repository, R3 stores data on their university server, which is protected by security standards, their laptop, and encrypted external hard drives, for so long as the contract outlining the terms of data use allow it. R3 never stores all the data in all these storage spaces, just some data in some of the storage spaces, and only for some of the time. There was a lengthy struggle before R3 could store data on university servers due to both technical and legal complications.

Although some publishers do run their own repository services, R1's employer does not, but they do provide some general resources and guidelines on where to start looking for a suitable repository, and the differences between storage and preservation. R1 confirmed that their employer prefers authors to deposit their data in a formal repository where it can be preserved, rather than storing it somewhere informal. The publisher wants researchers to use repositories that assign persistent identifiers, that have long-term preservation and funding solutions, and that utilise data equivalence services such as LOCKSS[8] and CLOCKSS[9]. R2 also noted that, while there is some use of formal repositories like Figshare in the Digital Humanities, researchers seem more likely to use services such as Humanities Commons or Zenodo to share their data.

In R1's opinion, "community owned data repositories are the best place for research data" because "where a community in a certain subject has created a space for that data to be held, they've set the standards" and are the curation experts. However, he noted that such services are not available in all subject areas (especially humanities subjects), so researchers must turn to broader repositories. R1 noted that their employer does not run its own repository because "publishers could never provide that gold-standard level of data support" that subject-specific communities can provide, and even if they tried, there would likely be pushback from data scientists and curators. R1 made it clear that their

[8] LOCKSS is "a widely-accepted best practice in the digital preservation field and more broadly for ensuring the persistence of digital information" (LOCKSS, n.d).
[9] CLOCKSS "provides a sustainable dark archive to ensure the long-term survival of web-based scholarly content" (CLOCKSS, n.d).

employer, as a company, “do not want to take ownership of researchers’ data”, because data “should [belong to] the researchers to do with what they want”.

4.1.7 Data citation and re-use

R1 noted the potential for data citations in the humanities specifically, due to the diverse research objects that researchers use. Data sharing includes ensuring functional links to datasets in journal articles, which requires technical work and systems integration (R1). R1 further noted that such technical infrastructure would allow for more accurate reporting in terms of how data are used and reused, and considered that such infrastructure could ultimately influence how researchers are acknowledged for interacting with data at different stages of their research, rather than just the publication stage.

R3 noted the difficulty of citing data if it cannot be assigned a persistent identifier in a repository, and confirmed there is a proposal underway for the assignment of persistent identifiers to archived web content in a repository, but that doing so is not straightforward since it involves collaboration with the International Organization for Standardization (ISO) and other organisations. Although it is difficult to cite archived web content without persistent identifiers, R3 noted that it is possible to publish a description of the data and how it was extracted in the form of reports, and placed in a repository accessible by others. R2 also emphasised the value of being able to link to data via persistent identifiers in published articles and books.

4.1.8 Publisher support and responsibility

R1 noted that, while it is the researcher’s responsibility to ensure their data are made available wherever possible and when required, publishers do have an important role to play in providing some guidance and developing infrastructure to support data sharing. As mentioned previously, R1 is adamant that publishers play a supportive rather than prescriptive role when it comes to researchers managing and sharing their data.

R1 confirmed that, currently, most guidance offered by their employer concerns publishing, sharing, and citing data. Apart from repository advice, the publisher provides templates that assist authors with completing a DAS. Both R2 and R1 spoke of their employer having offered some external-facing author and editor workshops and webinars on data sharing.

R1 sees one of his employer’s biggest responsibilities as being the installation of adequate technical infrastructure to accommodate links between published articles and datasets.

Another significant publisher responsibility concerns data planning through the development of a workflow that facilitates registered reports. Because registered reports allow for peer review of the research plan and design *before* the study is conducted, it is possible to change the plan before data collection has begun (R1). R1 clarified that any changes made to the data plan would not be prescribed by the publisher, but rather by expert peer reviewers. So, while the guidance itself would come from an academic in the subject area, the publisher's involvement concerns setting up the appropriate workflow to facilitate this process (R1). Registered reports would provide support during the data planning stage of research, thus re-balancing the publisher's current focus on data sharing and citation (R1).

R2 noted that although most publisher support currently addresses the data sharing and citation stages of the research lifecycle, publishers could try engage with and support researchers in the earlier stages: it could be valuable for publishers to provide guidance in presenting humanities researchers' methodologies or audit trails. R2 explained that "when it comes to […] giving credit for the work being done", such an audit trail would make it clear what work has been done without having to share the underlying data. This could increase the "transparency and visibility to a lot of the research process that currently goes underserved by publication processes" (R2).

R3 asserts that publishers should maintain a flexible approach to supporting data management and data sharing for WAS researchers, and that they should encourage it, but not impose strict requirements, mainly because of legal restrictions that researchers face in collecting, using, and sharing data. Flexibility is preferable so that authors can choose the most appropriate data sharing methods for their particular subject area (R3). R3 gave an example where a team of co-authors had shared some data in a repository and, when publishing a research article that used the data, they were required to deposit the data in a different repository, which proved a difficult and time consuming task.

In terms of the support currently offered by R1's employer to one of its internet studies journals, researchers are given the option to share data, but are not required to do so. Although the journal is still very new, having only launched a few years previous to the current study being conducted, only one author has shared their data so far. When asked if any training had been provided to the journal's researchers by the publisher, R3 mentioned there was an informative webinar to share information, but that this has not been a focus due to a lack of demand. R3 said the journal had not been approached by authors who want to share their data, so the publisher has not yet needed to prioritise training or workshops.

R3 responded positively when asked if WAS researchers could benefit from a publisher helpdesk that could advise on the specific legal complications of sharing data, saying that this could accommodate a broad range of subject areas, and that it would be a good resource for editors as well as researchers.

4.1.9 RDM and data sharing stakeholder collaboration

Stakeholders include any party that has an interest in and who might influence how research data are created, managed, used, stored, or shared. R1 mentioned that there are various groups discussing data sharing policies more broadly, how these are implemented by different subject areas, and that various stakeholders contribute to such groups, including publishers, repositories, and funders. R1's employer has formal relationships with organisations such as Figshare and Code Ocean, and less formal collaborations with organisations like DARIAH (R1). R2 and R1 both noted the establishment of a cross-publisher collaborative working group in 2020, which was installed as part of the STM initiative: 'Research Data Year'. The working group focuses on how best to support humanities data as publishers (R2, R1). One of the group's aims is to develop data-related case studies in the humanities to investigate how things have been done previously, which will inform how to do things in future (R2). R1 noted that, in future, they hope that a bank of such humanities-related case studies regarding data will be developed. The work being done through Research Data Year also involves collaborative work with stakeholders such as Crossref and COS (R1).

R1 noted that many future developments in the field will depend on how data management and sharing are evaluated and recognised – "if [data] is really going to become more central to the way that research happens, it needs to be acknowledged by all the stakeholders, including funders and researchers and institutions".

From a researcher perspective, R3 noted their involvement with WARCnet, and with a practical working group concerning WAS and research data management that seeks to map and offer solutions to the challenges faced when sharing web archive data across international borders.

4.1.10 Concluding remarks on interviews

Each interview lasted approximately 1.5 hours. R1 and R3 answered all questions asked, while R2 could not answer some questions relating to a journal, specifically whether researchers are encouraged to share their data at the point of article submission, and where researchers tend to store their data.

4.2 Questionnaire data

The questionnaire had a total of 31 responses. Seven were incomplete, but the answers that were given have been included in the data analysis. The questionnaire asked 49 questions, many of which were contingent on answers to preceding questions, so some questions were irrelevant to respondents, and not presented. Each question is addressed herein, and arranged in the order of the sub-sections in the questionnaire, which were guided by the JRDL.

Generally, questions attempted to ascertain detailed information about the nature of the data that WAS researchers tend to work with, the current data management and sharing practices of said researchers, and the challenges they experience in managing and sharing data. Questions also sought to elicit what support structures are available to WAS researchers regarding data management and sharing, and what support they would hope to receive in future, specifically from publishers. In most cases, respondents were given the option to select all multiple-choice answers that applied to them; where respondents were asked to choose only one option, this is noted in the following narrative.

The questionnaire data forms the quantitative part of the current study. Where relevant, comments on interesting correlations with the qualitative data are made. The full list of questions is available in Appendix D.

4.2.1 Part 1: Researcher information

Question 1 asked respondents to confirm their affiliations in order to assess their expertise and relevance to this study. This was the only demographic information collected. Twenty-five responses were collected, all of which confirm relevance to the study – either researchers were affiliated directly with internet, web studies, or Digital Humanities centres, or they were affiliated with humanities faculties, but conducting work in the field of WAS. Twenty-four respondents noted their geographical locations: USA (5), Canada (4), England (3), Ireland (3), Japan (1), Czech Republic (1), Denmark (1), Netherlands (1), Russia (1), Australia (1), no country stated (3).

Question 2 enquired whether respondents have published in an academic journal, book, or conference compilation in the last 18 months. 20 respondents (71.43%) answered ‘yes’, and 8 respondents (28.57%) answered ‘no’. This question was asked to ascertain whether respondents may have come across publishers’ data sharing policies.

4.2.2 Part 2: Attitudes toward RDM and data sharing in the field of Web Archive Studies

Questions 3 and 4 gauged the general attitude of WAS researchers to RDM and data sharing. Respondents were required to select only one answer for each question, the responses to which are presented in Figures 2 and 3. The vast majority of respondents thought RDM and data sharing to be very important for WAS. While this correlates with what interviewee R3 mentioned, R3 did note that data sharing is not always necessary for the field, which is perhaps reflective of the 20% of questionnaire respondents who were unsure of the importance of data sharing (see Figure 3).

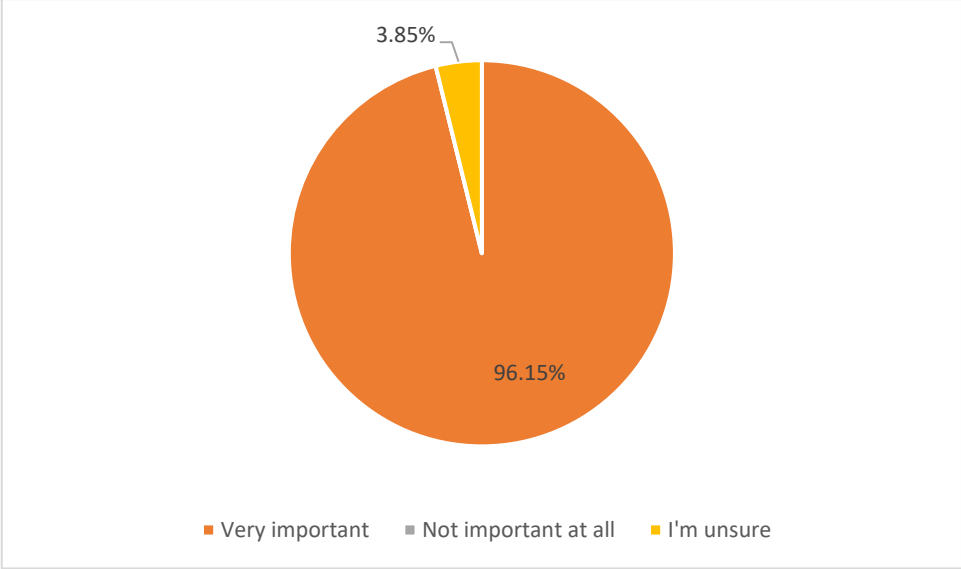

Figure 2. *Researchers' perceptions of the importance of RDM in Web Archive Studies (n=26).*

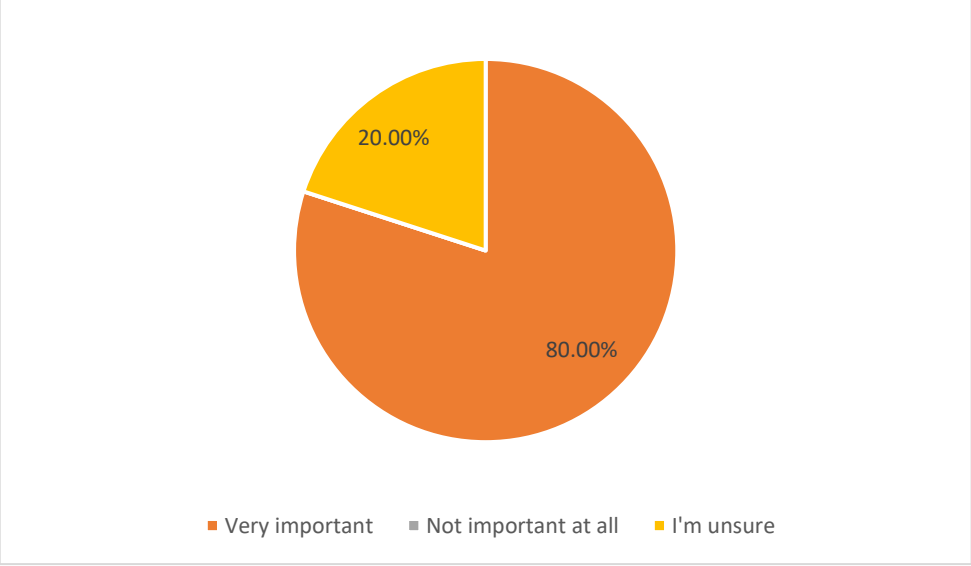

Figure 3. *Researchers' perceptions of the importance of data sharing in Web Archive Studies (n=25).*

4.2.3 Part 3: Types and sizes of data

Part 3 of the questionnaire gathered information on the kinds of data that WAS researchers tend to work with, as well as any specific challenges experienced regarding said data.

As shown in Table 2, most respondents (24; 92.31%) use the term 'data' to refer to their research, but the results indicate that WAS researchers also use other terms, such as those listed under 'Other'. This correlates with what interviewees R1, R2, and R3 said.

Table 2. *Terms used by Web Archive Studies researchers in relation to their research data (n=26).*

Answer Choices	No. of total responses	% of total responses
Data	24	92.31%
Research materials	5	19.23%
Other (please specify): ○ Resources (1) ○ Depends on the nature of the research (1) ○ Metadata (1) ○ Sources (1) ○ Depends on the audience – I do a lot of interdisciplinary work and public-facing writing, so I may clarify distinctions between data and research materials or explain why I'm using the term 'data' to audiences who might not value that term in particular contexts (1) ○ One respondent indicated they simply refer to specific format types (i.e. WARC files, HTML code, etc) (1)	6	23.08%

Results for Question 6, where respondents were asked to indicate the kinds of data they use, are depicted in Table 3 in order of most to least popular. Results reflect that WAS researchers use a diverse range of data, which correlates with what interviewees R1 and R2 said about the diversity of humanities data.

Table 3. *Data types used by Web Archive Studies researchers (n=26).*

Answer Choices	No. of total responses	% of total responses
Text files	22	84.62%
Images	22	84.62%
Archival metadata	18	69.23%
Publications (e.g. journal articles or books)	15	57.69%
Graphical images	14	53.85%
HTML code	13	50.00%
Numerical data (e.g. statistics)	10	38.46%
Audio files	9	34.62%
Other (please specify): ○ Executable software (1) ○ Audio-visual data (1) ○ Data from the archive sites (1) ○ WARC files (1) ○ 3D models (1) ○ XML-coded transcriptions of primary sources, digital surrogates for primary sources, documentation (1) ○ URLs (1) ○ Families of webpages (1)	8	30.77%
Geospatial data	7	26.92%
Field notes	6	23.08%
Crawl logs	4	15.38%

Regarding volumes of data (Question 7), 19 respondents (79.17%) selected 'gigabytes', 12 respondents (50%) selected 'megabytes', 7 respondents (29.17%) selected kilobytes, and 3 respondents (12.5%) selected terabytes, showing a wide range of data sizes being used.

Table 4 presents responses regarding the challenges associated with the type and size of research data used (Question 8) and shows that a slight majority of researchers say that the size of their raw datasets made computer processing a challenge.

Table 4. *Challenges experienced by Web Archive Studies researchers regarding data type and size (n=24).*

Answer Choices	No. of total responses	% of total responses
The raw data was too large in size for my computer to process	14	58.33%
The raw data was duplicated in certain instances	10	41.67%
Other (please specify): ○ Metadata for controlling and managing the data (1) ○ Data is not easily retrievable because cataloguers do not speak the language of the archive (Hawaiian) (1) ○ Variation in performance between multiple platforms and lack of technical documentation for emulation as a service (EaaS) (1) ○ Legal compliance of data provision and use (1) ○ Different types/formats of data, difficulties with processing data from home broadband connection, large datasets (1) ○ Predominance of data in English, limited access to web analytical tools (1) ○ Finding secure long-term storage with live availability and selecting formats that ensure long-term viability (1) ○ Incomplete data due to various issues such as crawling, storage, incompatibility with the modern software (1) ○ Diversity of data types (1)	10	41.67%
The original data set was too small to base a study on	5	20.83%
I have experienced no challenges	4	16.67%

4.2.4 Part 4: Size of researcher team

Questions in Part 4 of the questionnaire aimed to ascertain if WAS researchers tend to work collaboratively (with a research team), and what the associated challenges might be.

10 respondents (41.67%) stated that whether they work alone or in a team “depends”, while 7 respondents (29.17%) have worked alone, and 7 respondents have worked with a team (Question 9). If respondents answered ‘with a team’ or ‘it depends’, three conditional questions (Questions 10-12) were asked.

Most respondents who work in research teams stated their team size is between 1 and 10 members (Question 10). All responses are depicted in Table 5, which shows that some researchers work in considerably larger teams.

Table 5. *Size of research teams in the field of Web Archive Studies (n=18).*

Answer Choices	No. of total responses	% of total responses
1-10 research team members	17	94.44%
10-20 research team members	4	22.22%
20-30 research team members	2	11.11%
30-50 research team members	1	5.56%
Over 50 research team members	1	5.56%

As shown in Table 6, illustrating responses to Question 11, respondents mostly collaborate through email and digital meetings, with 17 respondents (94.44%) selecting each of these options.

Table 6. *How Web Archive Studies researchers collaborate with team members (n=18).*

Answer Choices	No. of total responses	% of total responses
Email	17	94.44%
Digital meetings	17	94.44%
Collaborative tools	16	88.89%
In person meetings	10	55.56%
Other (please specify): GitHub (2)	2	11.11%

The biggest challenge they seem to face in working with a research team is team member coordination (10; 66.67%) (Question 12). As can be seen in Table 7, one respondent highlighted a lack of funding as a challenge, which was also raised as a general concern for the humanities by interviewee R2. Another respondent highlighted that acknowledging individual contributions is a challenge, which was confirmed by interviewee R1, who stated that developments in terms of data management and sharing will only become a priority if researchers are properly credited for the work they do at different stages of their research, and acknowledgements given by all relevant stakeholders.

Table 7. *Challenges for Web Archive Studies researchers in working with multiple team members (n=15).*

Answer Choices	No. of total responses	% of total responses
Coordinating multiple team members	10	66.67%
Delegating work fairly	8	53.33%
Working across time zones	6	40.00%
The systems used to collaborate are not adequate for our purposes	5	33.33%
The size of the team	4	26.67%
Other (please specify): ○ Ensuring proper credit for individual contributions to research (1) ○ Lack of funding (1) ○ Gauging data literacy, establishing workflows or data creation or revision (1)	3	20.00%

4.2.5 Part 5: Data management planning

Part 5 of the questionnaire aimed to ascertain WAS researchers' current practices and challenges experienced regarding data management planning.

In response to Question 13, 15 respondents (60%) said that they had been required to complete a DMP, while 10 (40%) said they have not. Interviewee R3 also indicated never having been required to complete a DMP. If respondents answered 'yes', two conditional questions (Questions 14-15) were subsequently asked.

When asked why respondents had completed a DMP (Question 14), 10 (66.67%) chose both 'funder requirement' and 'researcher's choice', and 6 (40%) indicated having created DMPs due to an 'institutional requirement'. Responses to the second conditional question (Question 15) can be seen in Table 8, which shows that issues around tools and lack of knowledge about DMPs were experienced the most.

Table 8. *Challenges for Web Archive Studies researchers in creating a DMP (n=14).*

Answer Choices	No. of total responses	% of total responses
Issues with data management planning tools and systems	8	57.14%
Uncertainty about what to include in a DMP	7	50.00%
Unfamiliarity with the purpose of a DMP	3	21.43%
Other (please specify): ○ Metadata requirements and upload tools (1) ○ Multiple types of data collection, lack of guidance/literature for the management of archived web data extracted from web archives (1) ○ Lack of institutional infrastructure to support (1)	3	21.43%

If respondents answered 'no' to Question 13, a conditional, open-ended question (Question 16) enquired how they think a DMP might have helped their research in the past. Three respondents were unsure how a DMP would have helped, two stated that a DMP would have helped them to collect data, and one indicated that a DMP would have helped ensure the long-term survival of their data.

Interestingly, interviewee R2 stated that DMPs might not always be necessary for humanities research since 'data' planning would often be addressed in a literature review and reference list. As per Table 4, it is evident that there are WAS researchers using publications as 'data', though it is not clear whether they do so exclusively. Interviewee R3 also stated that a DMP would be difficult to maintain for WAS projects.

4.2.6 Part 6: Data collection and analysis

Part 6 of the questionnaire investigated data collection and analysis practices, and the associated challenges.

Question 17 asked whether respondents had used secondary or pre-existing data generated by third parties. Respondents were required to select only one answer – whether or not they had used pre-existing data – the results of which are presented in Figure 4.

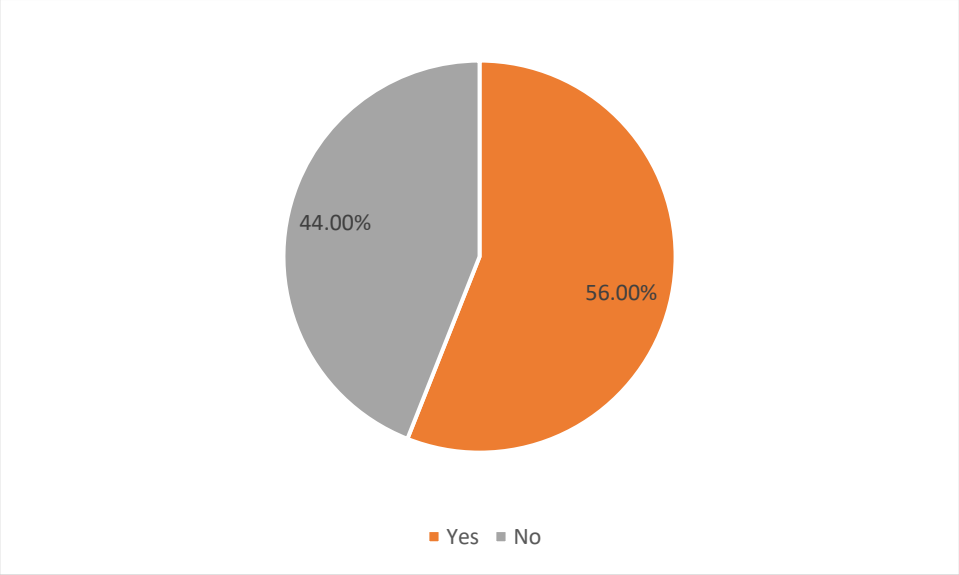

Figure 4. *Web Archive Studies researchers who have used pre-existing or secondary data (n=25).*

If respondents had used pre-existing data, they were asked three conditional questions (Questions 18-20) about finding data, getting permissions, and citing in publications. Most respondents (6; 42.86%) said they found data in a repository (as seen in Table 9). Likewise, most (11; 78.57%) said the data they had used was already licensed for re-use, so seeking permissions was unnecessary (see Table 10). Lastly, most respondents also indicated having subsequently cited the dataset in a publication (see Figure 5).

Table 9. *How Web Archive Studies researchers found pre-existing data (n=14).*

Answer Choices	No. of total responses	% of total responses
Found the data in a repository	6	42.86%
The data was shared via a file sharing service (e.g. Google Drive, iCloud)	4	28.57%
Other (please specify): ○ All of the above (1) ○ Included with Matlab package (1) ○ Data was harvested (1) ○ All of the above, and data retrieved from obsolete hardware (1)	4	28.57%

Table 10. *How Web Archive Studies researchers obtained permission to use pre-existing data (n=14).*

Answer Choices	No. of total responses	% of total responses
The dataset was licensed for re-use	11	78.57%
Obtained permission from the person or organisation who owns the data	7	50.00%
Other (please specify): ○ Harvested from open repository (1)	1	7.14%

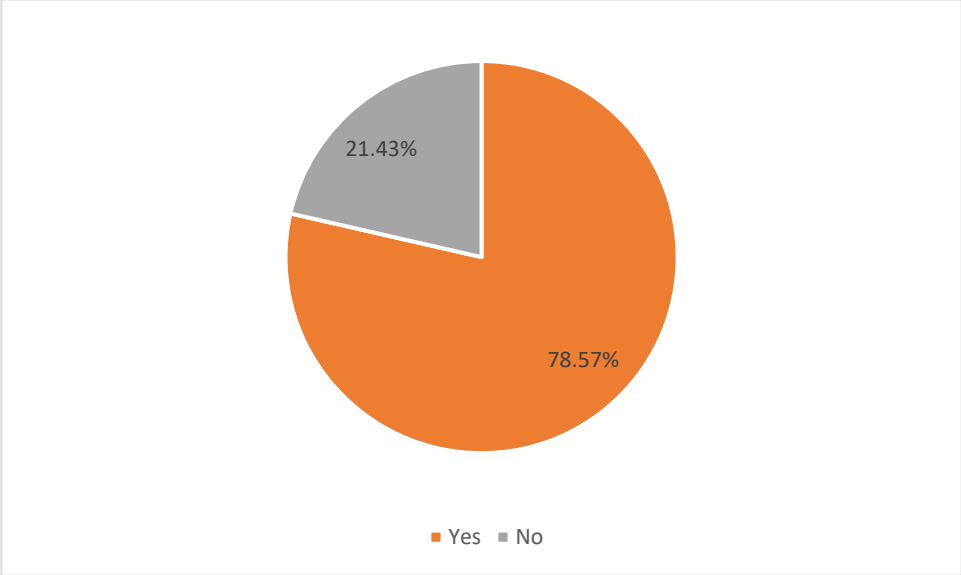

Figure 5. *Web Archive Studies researchers citing pre-existing data in a publication (n=14).*

All respondents were subsequently asked whether they had generated new data for their research (Question 21) – most of them had, as can be seen in Figure 6.

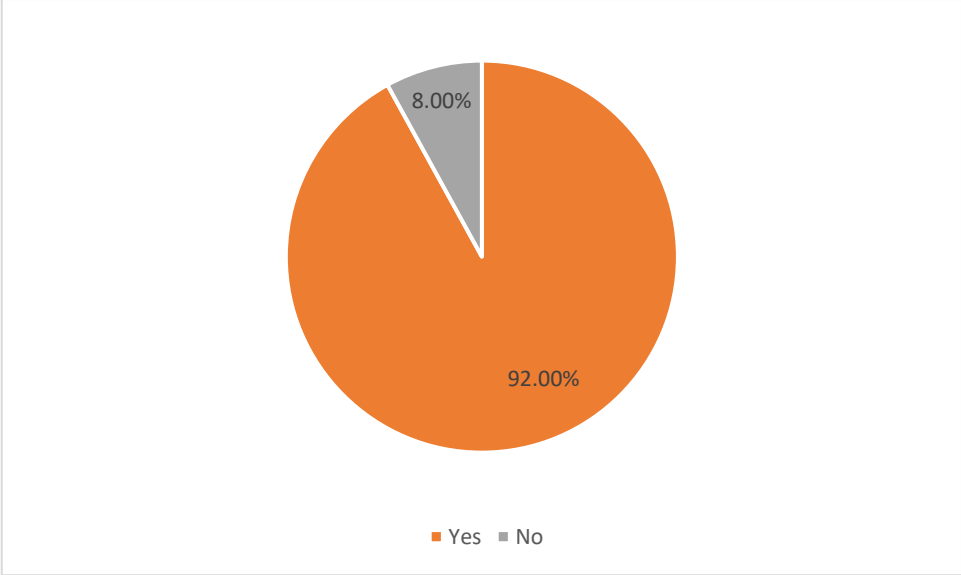

Figure 6. *Web Archive Studies researchers who have generated new data (n=25).*

New data require that file naming choices be made: dates, numbers, and places were the top three file naming conventions used by respondents in managing their data (Question 22), as shown in Table 11.

Table 11. *Files naming conventions used by Web Archive Studies researchers (n=24).*

Answer Choices	No. of total responses	% of total responses
Dates	19	79.17%
Numbers	16	66.67%
Places	12	50.00%
Initials	7	29.17%
Event	7	29.17%
Other (please specify): ○ A combination of naming conventions (1) ○ Names (1) ○ It's not always the same (1) ○ Project identifiers/codes (1) ○ ASCII – only lower case or camelCase, no punctuation, no spaces (1) ○ Names and titles (1)	6	25.00%
Comments	5	20.83%

In order to uncover whether WAS researchers tend toward pre-existing research methodologies (as suggested by the literature), respondents were asked whether they used a pre-existing method, process, workflow, or model to analyse their data (Question 23). As illustrated in Table 12 below, most

researchers had used a pre-existing method, indicating the value and usefulness of making methodologies available for re-use.

Table 12. *Web Archive Studies researchers use of pre-existing methods, processes, workflows, and models (n=19).*

Answer Choices	No. of total responses	% of total responses
Method	12	63.16%
Workflow	8	42.11%
Process	7	36.84%
Model	5	26.32%
Other (please specify): ○ None of the above (1) ○ We had to create our own at times (1) ○ Didn't understand the question (1)	3	15.79%

Question 24 asked respondents how they kept a record of their data collection and analysis methods, the results of which are outlined in Table 13, with the most popular being that researchers maintained version control and kept a research diary. This correlates with what interviewee R3 stated about the necessity of meticulous record keeping during the data collection and analysis stages of WAS research.

Table 13. *Web Archive Studies researchers' methods of documenting data collection and analysis (n=23).*

Answer Choices	No. of total responses	% of total responses
I maintained version control of all files.	19	82.61%
I kept a research diary or notebook.	14	60.87%
I kept a time log.	5	21.74%
I kept photographic records.	3	13.04%
I kept audio recordings.	3	13.04%
Other (please specify): ○ Methodology write-ups, workflow checklists, screenshots (1) ○ Shared Google Drive folder with relevant notes and documentation (1) ○ I maintained a spreadsheet of the names and contents of my other spreadsheets (1)	3	13.04%

Question 25 asked respondents what challenges they experienced with data collection or analysis equipment. The top responses, as shown in Table 14, include challenges relating to computer software and a lack of storage space. Two of the 'other' responses mentioned a lack of funding.

Table 14. *Web Archive Studies researchers' challenges with data collection and analysis equipment (n=24).*

Answer Choices	No. of total responses	% of total responses
Computer software	11	45.83%
Lack of storage space	11	45.83%
Computer hardware	8	33.33%
I experienced no challenges	7	29.17%
Internet speed	7	29.17%
Other (please specify): ○ Training team members in tracking their own processes and organisation of data (1) ○ Lack of funding and technical support (1) ○ Lack of funding and institutional hardware ○ Version control issues (1) ○ Problems with narrow gate of the web archive to download data (1)	5	20.83%

When asked in Question 26 whether they had experienced any other challenges regarding data collection and analysis, most respondents noted not having enough time (see Table 15). Two respondents mentioned a lack of funding, which is evidently a challenge during all stages of research. One respondent also noted a lack of established research methods, which again highlights the necessity of making previously used methodologies available to WAS researchers. One respondent noted a lack of guidance and literature on RDM for WAS, and another stated that there was a lack of data analysis and sharing guidance for humanities researchers from publishers specifically.

Table 15. *Web Archive Studies researchers challenges with data collection and analysis (n=22).*

Answer Choices	No. of total responses	% of total responses
I didn't have enough time for this stage in the research	13	59.09%
I struggled to get permission to use third-party data	6	27.27%
Other (please specify): ○ Metadata for control and management is very time consuming (1) ○ Lack of software accessibility (1) ○ Academic publications not being invested in data analysis and sharing for the humanities (1) ○ A lack of established research methods (1) ○ Lack of guidance and literature on RDM for web archive data (1) ○ Lack of funding (2)	6	27.27%
I have experienced no challenges	5	22.73%

4.2.7 Part 7: Data storage

Questions were asked to ascertain WAS researchers' current data storage practices, as well as the challenges they face during this stage of research.

Question 27 asked respondents (n=24) where they store their data. Most respondents (19; 79.17%) stated using their computer hard drives for this purpose, while 13 (54.17%) selected the cloud, and 11 (45.83%) indicated using external hard drives. Seven respondents selected 'other', four of whom mentioned an institutional repository or server.

When asked in Question 28 about long-term storage, the top three responses included the cloud, and external and internal hard drives. Responses can be seen in Table 16.

Table 16. *Methods Web Archive Studies researchers use for long-term data storage (n=25).*

Answer Choices	No. of total responses	% of total responses
The cloud (e.g. Google Drive, DropBox, or iCloud)	17	68.00%
External computer hard drive or flash drive	13	52.00%
Internal computer hard drive	10	40.00%
Other (please specify): ○ Institutional repository (1) ○ Online (1) ○ LTO tapes (1) ○ Encrypted external hard drive (1) ○ Institutional server (1) ○ Network storage (1) ○ Not possible to store because of GDPR (only derived results and code) (1) ○ Git and svn (1)	8	32.00%
Printed documents	6	24.00%

For Question 29, respondents were asked how much of their collected data are stored long-term. Exactly half (12) of the 24 respondents indicated storing 'all of it', and the other half 'some of it'.

When asked how long they usually intend to keep data once a study is completed (Question 30), most respondents (14; 58.33%) said longer than 10 years. Responses can be seen in Table 17.

Table 17. *How long Web Archive Studies researchers keep their data for (n=24).*

Answer Choices	No. of total responses	% of total responses
Longer than 10 years	14	58.33%
Between 2-5 years	3	12.50%
Between 5-10 years	3	12.50%
Less than 1 year	2	8.33%
Between 1-2 years	2	8.33%

Question 31 asked respondents what challenges they experienced regarding data storage. As seen in Table 18, storage space is the most frequent challenge, with funding and IP also considered to be notable challenges by over a third of respondents. The latter challenges were raised as concerns across both the questionnaire and interviews. Six respondents (24%) selected 'other', with varying reasons given, and no answer being mentioned more than once.

Table 18. *Challenges that Web Archive Studies researchers face regarding the storage of data (n=25).*

Answer Choices	No. of total responses	% of total responses
Storage space challenges	13	52.00%
Funding challenges	11	44.00%
Intellectual property challenges	9	36.00%
Privacy challenges	6	24.00%
Other (please specify): ○ Metadata challenges (1) ○ Format change challenges (1) ○ Logistical challenges (keeping track of external drives, shared cloud folders, etc) (1) ○ Researchers moving institutions (1) ○ I cannot store my data due to legal restrictions (1) ○ Experienced no challenges – DMP already outlined how the data would be shared based on permissions granted (1)	6	24.00%
Security challenges	4	16.00%
I have experienced no challenges	3	12.00%

4.2.8 Part 8: Data sharing

The following questions enquired after WAS researchers' current data sharing practices and the related challenges.

In response to Question 32 (n=25), the majority (15; 60%) indicated having shared their data in an institutional or subject repository (in contrast to interviewee R2's opinions), and were subsequently

asked which repository they used (Question 33). Although few answered this question, most respondents said they had shared their data in an institutional repository. All responses can be seen in Table 19.

Table 19. *Repositories used by Web Archive Studies researchers (n=9).*

Answer Choices	No. of total responses	% of total responses
Institutional repository	4	44.44%
Docubase	1	11.11%
GitHub	1	11.11%
FigShare	1	11.11%
National repositories (CORA.ucc.ie and DRI.ie)	1	11.11%
Zenodo	1	11.11%

All respondents were then asked about the alternative ways they have shared their data with others (Question 34), the responses to which are outlined in Table 20. The top three responses include file sharing services, email, and social media.

Table 20. *Ways that Web Archive Studies researchers have shared their data, apart from repositories (n=19).*

Answer Choices	No. of total responses	% of total responses
File sharing services (e.g. Google Drive, iCloud)	15	78.95%
Email	10	52.63%
Social media (e.g. Facebook, Twitter, WhatsApp)	7	36.84%
File transfer services (e.g. WeTransfer)	5	26.32%
Other (please specify): ○ Web pages ○ Physical HDD transfer ○ GitHub ○ Blog publications	5	26.32%

Question 35 asked respondents if their data were assigned persistent identifiers when shared. Eleven respondents (52.38%) answered 'no', and 10 (47.62%) answered 'yes'. Interviewee R3 also noted not having used persistent identifiers, but that work is underway to develop a national repository specifically for web archive data, which would allow for this.

Question 36 asked respondents what kind of information they included in the metadata describing their data. As outlined in Table 21, the top five responses include description, author, title, keywords, and subject. Only 5 respondents (21.74%) said that no metadata were generated.

Table 21. *Metadata that Web Archive Studies researchers include with their data (n=23).*

Answer Choices	No. of total responses	% of total responses
Author	16	69.57%
Description	16	69.57%
Title	14	60.87%
Keywords	14	60.87%
Subject	13	56.52%
Publisher	9	39.13%
No metadata was generated	5	21.74%

Question 37 asked respondents what challenges they experienced regarding sharing their data (see Table 22), with top challenges concerning 'funding' and 'storage space', both options being selected by eight respondents each (40%). 'Security' and 'privacy' challenges were selected by five respondents each (25%), with four respondents (20%) selecting 'intellectual property' challenges. These complement what interviewee R3 noted – that the legal challenges have been the most difficult regarding data sharing for humanities researchers, which was echoed by R1 and R2 too.

Table 22. *Challenges to sharing data for Web Archive Studies researchers (n=20).*

Answer Choices	No. of total responses	% of total responses
Funding challenges	8	40.00%
Storage space challenges	8	40.00%
I have experienced no challenges	6	30.00%
Security challenges	5	25.00%
Privacy challenges	5	25.00%
Intellectual property challenges	4	20.00%
Other (please specify): ○ Metadata challenges (1) ○ Software decay (for web-based projects) (1) ○ Bandwidth latency (1) ○ I envisage no data sharing problems as DMP was designed to ensure data collection/sharing were combined as final output (1)	4	20.00%

4.2.9 Part 9: Data publication

Part 9 of the questionnaire concerned the data management and sharing requirements placed on WAS researchers when publishing their research. For all three questions (Questions 38-40), just 'yes' or 'no' answers were required in response to whether a publisher has ever required them to submit a DAS,

adhere to a data sharing policy, or to submit a DMP. Although 24 respondents answered these questions, only 20 respondents (71.43%) have published in the last 18 months, as per Question 2.

Responses, which were mostly in the negative, can be seen in Figures 7-9.

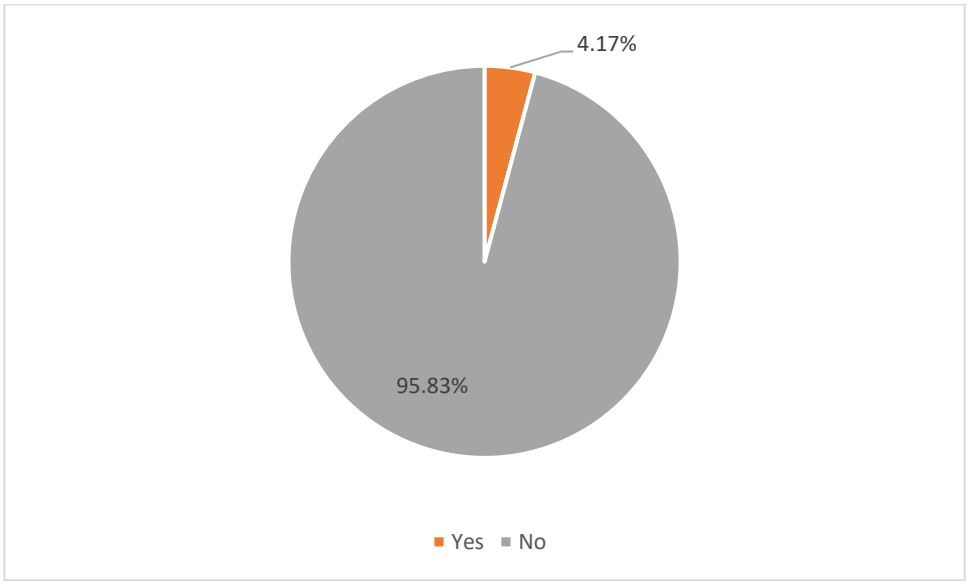

Figure 7. *Web Archive Studies researchers who have been required by a publisher to submit a Data Availability Statement when publishing research (n=24).*

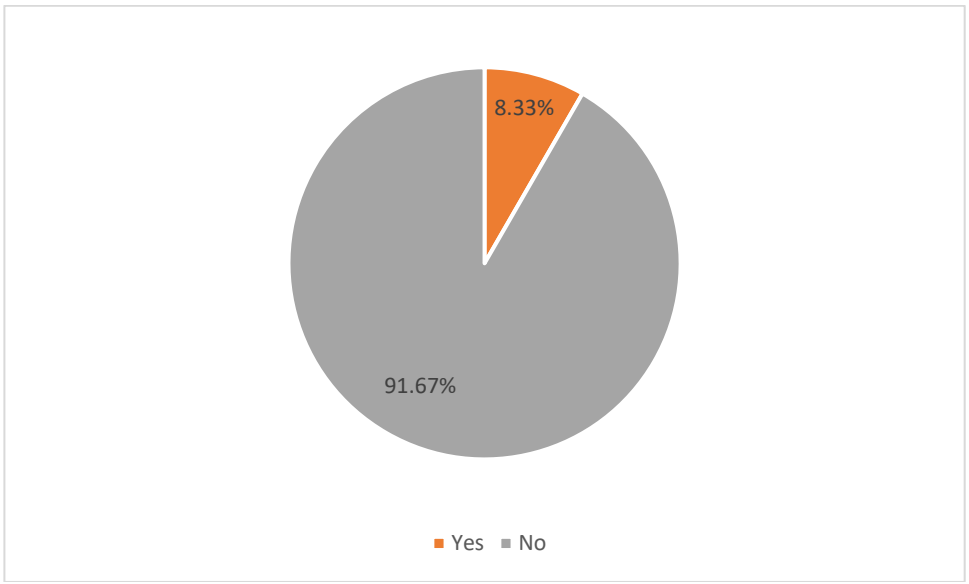

Figure 8. *Web Archive Studies researchers who have been required by a publisher to adhere to a data sharing policy when publishing research (n=24).*

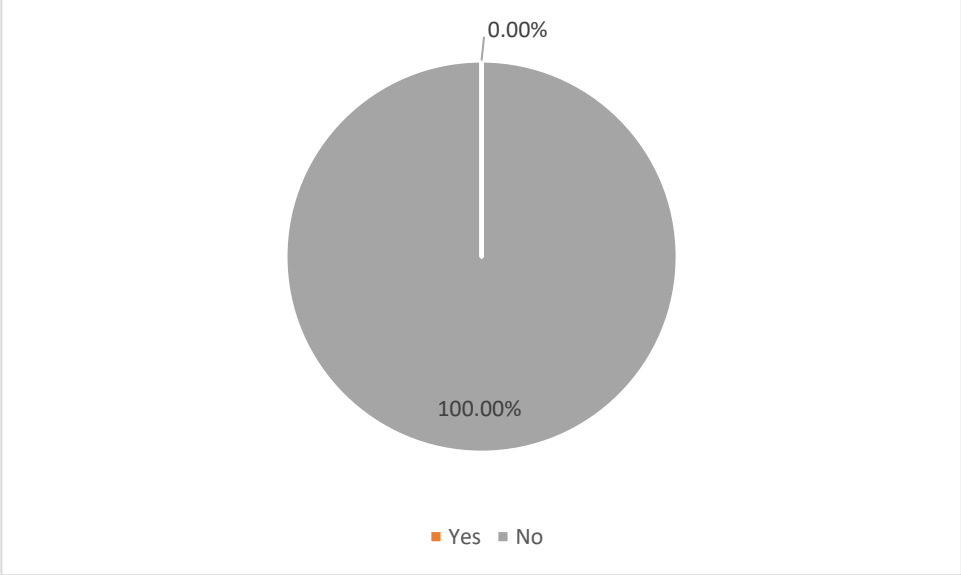

Figure 9. *Web Archive Studies researchers who have been required by a publisher to submit a DMP when publishing research (n=24).*

4.2.10 Part 10: Data re-use

Part 10 of the questionnaire was concerned with how WAS researchers re-use data.

Question 41 asked respondents how long they take to re-use their data after collecting it, with the option 'I don't re-use my data' being selected by half of the 12 respondents; the remaining six indicated that their data are re-used between 1 and 2 years after the being collected.

Question 42 asked respondents if their data had ever been used by anyone else. Answers were more or less evenly spread across the categories of 'yes' (8 respondents), 'no' (9), and 'I don't know' (7).

4.2.11 Part 11: Support for researchers

The following questions addressed the levels of support WAS researchers receive, and from which stakeholders, regarding data management and sharing.

In response to Question 43, all respondents (n=25) indicated having never received training from an academic journal or publisher to assist them with managing or sharing their data. Since no respondents answered 'yes' to the latter, the conditional question (Question 44), which asked what kind of training was received, was not presented.

Question 45 asked respondents whether they had ever been provided with any resources or guidelines by academic journal editors or publishers that assisted them with managing and sharing their data. 23 respondents (92%) stated 'no', and two (8%) stated 'yes'. Interviewee R1 similarly stated that

publisher guidance specific to the humanities and humanities data types does not yet exist, but which is something publishers are working towards providing. This also correlates with some of the answers to Question 26, where one respondent stated that publishers are not invested in data analysis and sharing for the humanities, and another noted a lack of guidance and literature on RDM for web archive data.

The two respondents who answered 'yes' to the latter question were subsequently asked what kinds of guidelines were provided to them by academic journal editors or publishers (Question 46). One respondent indicated 'general data sharing guidelines', and one said they were encouraged to upload datasets linked to a journal article.

Question 47 asked respondents what additional support from academic journals and publishers would be most useful to them as WAS researchers, the responses to which can be seen in Table 23. The top responses included publisher collaboration, providing repository guidance, and providing written guidance on the benefits of RDM and data sharing. Nine respondents (37.5%) mentioned that publishers should implement stricter data sharing policies, contrasting interviewee R3's comment that publishers should remain as flexible as possible regarding data sharing policy requirements. Five respondents selected 'other', one of whom said that publishers should be part of the larger conversation around data management, but that open repositories should remain the sole custodians of data. The latter was highlighted by interviewee R1, who stated that publishers should rather provide guidance on how researchers can manage and share data should they want to do so. One respondent also mentioned that data sharing is often not possible because of copyright and GDPR policy, a relatable problem in interviewee R3's experience of working in the field of WAS.

Table 23. *Support that Web Archive Studies researchers would find most helpful from publishers (n=24).*

Answer Choices	No. of total responses	% of total responses
I would like to see publishers collaborating more with institutions and libraries to offer more support to researchers in managing and sharing their data.	18	75.00%
I would like to receive information on which repositories I should use for my subject area.	17	70.83%
I would like to receive written information on the benefits of Research Data Management and data sharing (e.g. a handbook, online toolkit resources).	11	45.83%
I would like to receive Research Data Management training from a publisher.	9	37.50%
I think that publishers should introduce stricter Data Sharing Policies to ensure that researchers in Web Archive Studies are making their data available to others.	9	37.50%
I would like publishers to be more involved in Research Data Management and data sharing from an early stage in the research.	7	29.17%
I would like to receive data sharing training from a publisher.	6	25.00%
Other (please specify): ○ Institutional support is needed (1) ○ Security measures needed to honour tribal or native community or cultural/traditional rules for sharing information and protecting it Publishers and institutions need to have living relationships with the people who belong to the data (1) ○ Data sharing often not possible because of GDPR and copyright (1) ○ Publishers should be part of larger conversations about data management, but the prioritisation should be on open access repositories. Publishers should not retain exclusive or proprietary rights to data. (1) ○ Publishers are not the appropriate agents for data management and sharing. Open access repositories are more appropriate, and these should not be owned or managed by publishers. (1)	5	20.83%

When asked in Question 48 what support they have received from their institutions to assist them with data management and sharing, 10 respondents (41.67%) mentioned having been made aware of repositories where they can share their data. Nine (37.5%) noted having not received training from their institution, and eight (33.33%) said they had received training on RDM, and on how to complete a DMP. Four (16.67%) mentioned having received training on data sharing and four selected 'other', with no single answer being mentioned more than once. One response was significant though – a respondent said they had taught themselves about RDM and data sharing by attending webinars and reading books on the subject.

4.2.12 Part 12: Final comments

Question 49 asked respondents if they had any final comments regarding data management and sharing, and a text box was provided for answers. One respondent said they appreciated the current study being done on RDM in the field of WAS. One respondent noted that it would be helpful to clarify what falls into the category of 'data', since they include all project outputs as data. One respondent noted that data stewardship and infrastructure is required at both local and national levels, and that funders need to support open and FAIR data.

4.2.13 Concluding remarks on questionnaire

Although the self-selected sample size of researchers was small, the fact that most respondents answered all the questions, despite the questionnaire being long, shows an interest in engaging with RDM and data sharing in WAS.

4.3 Chapter conclusion

This chapter presented data gathered from three interviews and a questionnaire, where responses include perspectives from both publishers and WAS researchers. The data presented addressed the research objectives, namely to provide details on the nature of data and the current RDM and data sharing practices of researchers in the field of WAS, and to discover the ways in which publishers are currently engaging with and supporting WAS researchers with their RDM and data sharing. A discussion of the data will follow in Chapter 5.

Chapter 5: Discussion, recommendations, and conclusion

This chapter discusses the key findings from Chapter 4, showing their relevance to the research objectives and comparing them to the findings of others outlined in the literature review. The chapter also makes some recommendations based on the conclusions drawn.

5.1 Data in Web Archiving Studies

One of the study's research objectives was to provide details on the nature of data in the field of WAS, and the results produced three key findings.

First, the results illustrate a multitude of data types that WAS researchers work with, confirming previous literature (Gómez, Méndez & Hernández-Pérez, 2016:547; Schöch, 2013). This study has found that data are diverse not only within the humanities, but also within WAS.

Second, for the most part, 'data' was found to be a term that can be used in the humanities, but that not all humanities researchers tend to do so. While many WAS researchers do identify with the term 'data', many also identify with other terms, and some do not identify with the term 'data' at all, as confirmed by both Henry (2014:347) and Schöch (2013). DARIAH (n.d.a) also confirms that many sources understood as 'data' in other subject areas are referred to differently in the humanities. Jones, Grant, and Hrynaszkiewicz (2019:1) and Wyatt (2019) both note 'data' as being an inadequate term for capturing the complexity of the research materials and sources that humanities researchers use. Interviewee R2's point that a term other than 'data' might be needed for the humanities is similar to Drucker's (2011) proposal – that there is a need to reformulate the term 'data' for the humanities.

Third, much of the data that WAS researchers in this study tend to use is pre-existing or secondary, which is a common observation within humanities and Digital Humanities research (for example, Poole & Garwood, 2019:2; Gómez, Méndez & Hernández-Pérez, 2016:547). While WAS researchers also generate their own data, the secondary nature of data often causes challenges related to IP and third-party ownership.

5.2 Strengths, weaknesses, and challenges of Web Archive Studies researchers' RDM and data sharing practices

Other research objectives concerned the data management and sharing practices of WAS researchers, and the related challenges and limitations experienced. To this end, this section discusses the key strengths and weaknesses of WAS researchers' data management and sharing practices as per the

research findings, as well as the key challenges experienced by them in managing and sharing their data. The strengths, weaknesses, and challenges all relate to the various stages of the JRDL.

5.2.1 Strengths of Web Archive Studies researchers' data management and sharing practices

The main data management related strength concerns research methodology. It is evident that most of the WAS researchers who participated in this study keep meticulous records of their research methods (and understand the necessity of doing so) during the data collection and analysis stages of research. This includes assigning metadata (albeit sometimes a challenge), using file naming conventions (mostly in the form of dates, numbers, and places), and employing research diaries and notebooks to maintain an audit trail of how the data were treated and interpreted. It is also clear that many participants used pre-existing methods and processes.

These findings indicate that researchers in the field of WAS see recording methodologies as integral to their work, highlighting the value of re-usable methodologies for the subject area. Brügger, Laursen & Nielsen (2019:124) and Schroeder and Brügger (2017:9) note the direct importance of keeping a clear record of the methods and processes used to create a corpus of web archive data. This also correlates with what Shenton (2004:72), Carcary (2009:11), and Bowen (2009:305) assert regarding audit trails being extremely valuable, especially for qualitative studies, because they contribute to the transparency and trustworthiness of a study. If it is not possible to share data, it is usually possible to share documentation showing how the data were collected and used.

In addition to keeping detailed records of their research methodologies, WAS researchers have experience filling out DMPs, for reasons other than a publisher requiring it, one being that it was the researcher's choice. One respondent mentioned how useful a DMP was to their data management and sharing efforts. WAS researchers store and back up either all or some of their data (some respondents storing it in multiple locations), and store at least *some* of their data long-term.

Of WAS researchers who had sourced and used pre-existing data before, most not only ensured that the data were licensed appropriately for re-use, but also cited the data in a research publication, showing, presumably, that they appreciate others sharing data. WAS researchers using secondary data implies that these researchers benefit from data sharing, although there is little evidence showing such data are shared by other WAS researchers. Most WAS researchers indicating that they generate new data shows there is potential for robust RDM processes from the start of research project.

There is evidence that WAS researchers *do* share data within their research community, even if often done so informally (for example, via email). WAS researchers also often work in teams, implying more

sharing potential, although this entails extra challenges. The majority of WAS researchers have shared their data in repositories, indicating a general ability and willingness to do so, even if interviewee R3 could not.

Such practices show a fairly advanced grasp of some important aspects of RDM, such as finding (and re-using), citing, organising, and even sharing data. This correlates with Borgman (2012:1061) and Munoz and Renear's (2011) statements regarding Digital Humanities researchers being relatively engaged with data management. Findings also corroborate Poole (2013), Wilms et al. (2019:28), and Poole and Garwood's (2019:10) assertions that RDM is of great importance to Digital Humanities research, and bodes well for WAS researchers' future best practices (if there is adequate support in place).

5.2.2 Weaknesses of Web Archive Studies researchers' data management and sharing practices

One weakness found relates to WAS researchers not having a clear definition of what data is, and not always seeing the materials they work with as 'data'. Evidently, this sometimes results in researchers not seeing data management as entirely relevant to them, even while seeing its importance generally. Such a lack of clarity when defining data in the broader humanities is noted by Borgman (2012:1060) and Beagrie (2019:8).

While WAS researchers do share their data with team members informally, they do not always share it in formal repositories (with no extant repository dedicated to web archive data). Doing so would ensure secure preservation and the use of persistent identifiers, which are important for findability and citation purposes. The latter indicates that data sharing is usually not mandated by funders, institutions, or publishers, despite respondents saying that funders do mandate DMPs. That said, evidently not all subject areas should use the same RDM and data sharing approaches, and mandates that work well for the sciences might not be suitable for WAS or other humanities fields.

A weakness regarding data sharing policies can be seen in the example of R1's employer rolling out a basic data sharing policy encouraging data sharing in their internet studies journal, but only one author having actually shared their data. This indicates an extremely low sharing rate, even when considering that the journal had been running for less than five years as of 2020. There is undoubtedly a gap when comparing WAS researchers' practices with those encouraged by the JRDL.

A further weakness regarding WAS researchers is the lack of an available repository specialising in web archive data that would allow data to be preserved long-term, also impacting researchers' abilities to

share data. This is another significant gap when comparing WAS practices with those encouraged by the Jisc lifecycle.

Other weaknesses when comparing study findings to lifecycle best practices include: (1) about a fifth of WAS researchers do not generate any metadata that they are aware of, which could compromise the ability to find and search for data after it has been stored; (2) the majority of WAS researchers do not reuse their data (not necessarily a weakness but this would depend on the research being done); (3) there is a clear uncertainty around creating DMPs, meaning WAS researchers might not see their value; and (4) many WAS researchers store data on their computer hard drives and in the cloud, and might not necessarily store their data in other locations, which is not a sound storage strategy.

Many of these weaknesses, however, result from the challenges described below.

5.2.3 Challenges for managing and sharing data in Web Archive Studies

A main challenge to WAS researchers trying to manage and share their data relates to legal restrictions, often involving either privacy or third-party ownership issues, sometimes preventing researchers from sharing their data with others. Difficulties in obtaining third-party permissions for data re-use has been highlighted as a common challenge for the broader humanities by researchers such as Borgman (2009:17,20) and Gómez, Méndez and Hernández-Pérez (2016:547). Such challenges affect nearly all stages of RDM.

Legal issues are sometimes complicated due to different regions having different legal restrictions regarding how data may be accessed and used – a point made during the interviews. If some researchers are better able to share their data than others, there will be implications for international funder and publisher data sharing policies, namely that they cannot be too rigid in their requirements. This was also touched on in the interviews, where it was suggested that publishers be as flexible as possible regarding data sharing requirements.

Apart from legal restrictions, it is also evident that a lack of funding is a challenge for WAS researchers' RDM practices during the data storage and sharing stages. For example, a lack of funding means that WAS researchers are not forced by funders to write DMPs, to store data long-term, or to share data in specific repositories. Despite a general interest in data curation within the field of Digital Humanities, there is still a lack of RDM instruction provided to researchers in this field (Dressel, 2017:5), and a lack of funding (including funder mandates to share data) is a likely reason for this.

The study found that a lack of guidance, instruction, and training regarding appropriate management and data sharing was also a large challenge for WAS researchers, adding to the fact that most extant publisher advice and guidance is more science focused. The implication here is that if there is a lack of guidance for WAS researchers, they cannot be expected to comply with policies, even if those policies encourage RDM best practice and data sharing.

Web archive data extraction can also be a challenge for WAS researchers due to the complex and varying ways that the data are archived. It's important to note R3's comment that if their research team had known that data extraction (and the legal restrictions) would be so challenging, they might not have initiated the project.

Two other key challenges concerned storage space and the size of raw data for computer processing, hinting at Poole and Garwood's (2019:5) finding that Digital Humanities researchers are often faced with difficulties regarding the technical requirements for storage.

Most study participants indicated having worked in a research team at some point, showing WAS to be a collaborative field, and thus reflecting Poole and Garwood (2019:1) and Borgman's (2009:20) assertion of the Digital Humanities as a collaborative field. However, some of the key challenges faced by researchers regarding teamwork include time zone issues, and difficulties with task delegation: managing large teams in the Digital Humanities is a challenge (Poole & Garwood, 2019:9). While also proving a challenge for WAS researchers, teamwork is also, however, a potential strength in that there are more resources for RDM, and more potential for sharing data.

WAS is an emerging field, meaning there are no RDM standards, procedures, or methods that researchers can refer to. Interestingly, when asked what challenges they had encountered during data collection and analysis stages in WAS, one researcher noted a lack of established research methods for the subject area. Although only one respondent mentioned this directly, it does highlight the importance of encouraging transparency and methods sharing in the field.

Disincentives are also challenges, in that they bar researchers from sharing their data, even if the data are managed and stored well.

Another challenge concerns the complex question of which parties (researchers, project managers, institutional libraries, funder, or publishers) should be responsible for RDM and data sharing (Poole & Garwood, 2019:8). According to the results herein, publishers should not be the custodians of data, but they do have an important role to play in providing guidance, and developing policies and workflows.

5.3 Publishers' support for data management and sharing in WAS

Many of the findings about publisher support relate to the following two points made earlier in the study: first, that RDM and data sharing infrastructure is far less developed in the humanities than in the sciences due to less funding and fewer data management and sharing requirements (Borgman, 2012:20; Poole, 2013; Poole and Garwood, 2019:4); and second, that there is a lack of humanities-related guidance offered to researchers by publishers, and no WAS-specific guidance at all.

Given that data management and sharing is more developed in the sciences than in the humanities, it follows that publishers' current support, guidelines, and resources offered to researchers have been developed in line with the needs and requirements of scientific areas of study. The latter was confirmed in the interviews, and most WAS researchers noted a lack of investment and, thus, lack of guidance from publishers in terms of RDM (though publishers were not specified, and a few researchers did receive advice). This study found an overall lack of publisher advice, training, guidance, and support for RDM and data sharing throughout WAS, although this is something the publisher employing R1 is actively working toward, in collaboration with other publishers and stakeholders. It is relevant to mention one participant's opinion that academic publications are not invested in data analysis and sharing for the humanities, while another noted a lack of guidance and literature on RDM for web archive data specifically. Some researchers would like to receive advice on which repositories to use – a point being addressed by R1's employer, but not specifically for the humanities. The need for WAS advice, guidance, and support is also present in researchers' desires for stricter data sharing policies, even though this may prove to be limiting. Most WAS researchers have not been asked for DASs or DMPs, or to adhere to data sharing policies.

One respondent did call for publishers to engage more with other stakeholders (particularly institutions and libraries) in support of WAS research data management and sharing. Simms et al. (2016) also call for more collaboration between stakeholders. Although there is currently limited collaboration evident in the literature between publishers and other organisations, this study discovered work being done between the R1's employer, and organisations such as DARIAH, RDA, and the STM Association. This publisher was fundamental in installing a cross-publisher working group in 2020 dedicated to the discussion and development of publisher policies and support for humanities researchers.

There is also work being done regarding the collaborative development of publishers' data sharing policies. Publishers are rolling out data sharing policies to their journals and are beginning collaborations for standardising such policies. Humanities journals, for the most part, seem to use

flexible policies that encourage, but do not require data sharing from researchers. It is also apparent that stricter policies are often not suitable for WAS and the broader humanities, where there are issues with IP or privacy that bar the sharing of data due to its qualitative and secondary nature. This correlates with the finding that most WAS researchers have never been required by a publisher to adhere to a data sharing policy, or to submit a DAS or a DMP when publishing their research in a journal or book.

There has, evidently, been limited training offered to researchers by publishers in the way of RDM or data sharing, and none for WAS specifically or the broader humanities so far. This is despite desires on the part of WAS researchers for such training. However, R1's publisher has been investing time and resources into building adequate workflows to accommodate registered reports and data citation links – two important aspects of data sharing infrastructure that can be applied to data in any subject area, including the humanities.

5.4 Recommendations for publishers

Drawing on the findings related to the study objectives, this section will propose ways in which publishers might better support WAS researchers in managing and sharing their data.

The first recommendation is for publishers to acknowledge the diversity of data types in WAS, and in the broader humanities and Digital Humanities, and that the term 'data' might not be an appropriate blanket term. Applying another blanket term would not solve the problem, however. The recommendation is therefore that publishers (and, by default, other stakeholders), to embrace this diversity of 'data' and reflect this in any guidance, advice, or resources made available to WAS researchers regarding data management, sharing, or citing.

Although there is existing guidance regarding data management and sharing for the sciences, this study does not recommend that such guidance be applied directly to WAS or other humanities subject areas. Given that the nature of humanities data are so diverse, it follows that extant approaches to RDM and data sharing for the sciences would not always work for WAS or humanities data types. This is in large part due to the secondary nature of much WAS and humanities data, often introducing legal restrictions to data use and dissemination. This study therefore recommends that publishers develop guidance specific to WAS researchers' data management and sharing needs.

This study also recommends that publishers focus on the sharing of methodological processes, audit trails, and research instruments, rather than the sharing of data, for WAS and other humanities subjects. This promotes transparency in subject areas for which data sharing is often not possible due

to legal restrictions. A number of extant repositories, such as Dryad or Figshare, could facilitate the stewardship of such methodological materials and instruments, including the assignment of persistent identifiers, meaning a completely new archival system would not be necessary, and publishers could remain uninvolved. Publishers could duplicate existing workflows to facilitate the links between published articles and shared methodologies and instruments, which potentially means less investment regarding system development and integration.

Publishers could adapt current journal data sharing policies to require the sharing of research methodology materials and instruments for humanities subjects, and ensure compliance, since there would be fewer legal restrictions limiting researchers' ability to share said material with others. Kvalheim and Kvamme (2014:7) state that formal policies are necessary to normalise RDM practices in the humanities. To this end, the policies could require researchers to submit a Methods Availability Statement, which could function similarly to current DASs that link to information in a repository. WAS and humanities researchers are generally accustomed to maintaining records of how their research was conducted, and policies requiring this would likely come as less of a shock than requirements to share data, and thus be more readily adhered to.

Publishers should develop guidelines in support of such policies, providing information on the differences between methodologies and data, the importance of sharing methodological processes and instruments for transparency, and provide details on the kinds of methodological materials and instruments that could be shared (for example, a research diary or logbook, annotations, a code book or matrix template, survey questionnaires, and interview questions). Such guidance could also advise on the general or subject repositories most suitable for preserving and making humanities methodological materials accessible. The guidance could still address data, and link to existing guidance on data sharing, but a more flexible stance on data sharing, with the emphasis on sharing the methodologies used to collect the data, is preferable. Importantly, such guidance should contain language already familiar to WAS researchers (and broader humanities researchers), for example, without a reliance on the term 'data'. Such guidelines might draw on and/or adapt the CESSDA Data Management Expert Guide (CESSDA, 2020), which is licensed for re-use. Additionally, publishers could offer training around policies and what the expectations are for WAS researchers. Indeed, Munoz and Renear (2011) state that, in order to promote sound RDM practices in the humanities, researchers need skills development, training, and education. Buddenbohm et al. (2016:32) and Simms et al. (2016) also note the importance of RDM and data sharing training for researchers in the humanities. And, ultimately, publisher training should ideally complement training provided by other stakeholders.

In order to implement policies that require WAS researchers to share their methodological processes and instruments, and in order to develop supporting guidelines, publishers will need to collaborate with other stakeholders such as libraries, institutions, funders, and repositories to ensure consistent messaging around the importance of research methods transparency in the humanities. This suggestion is supported in literature by Beagrie (2019:14) and Simms et al. (2016:32). Publishers should continue to collaborate with other publishers in developing standard approaches to supporting WAS researchers and those in the broader humanities and Digital Humanities, and to avoid divergent messaging.

Another recommendation is for publishers to install an internal helpdesk to assist researchers with legal advice and navigating data sharing challenges specific to WAS, and other humanities subjects that use secondary data. Publishers would be in a position to do so if they are already collaborating with and aligning their policies and guidance with those of other stakeholders.

Although not a recommendation for publishers, there is need for a repository dedicated to archived web data that can house extracted corpora in the long-term, assign persistent identifiers that identify the data as web archive data, and allow for different levels of security and access management. Although there was a call by interviewee R3 for such a repository on a national level, it might make more sense for a repository to serve WAS researchers internationally.

5.5 Future studies

Given that the sample size of researcher participants was small for the current study, a future study could conduct a purely qualitative case study with multiple in-depth interviews with both publishers and researchers. Alternatively, a similar mixed methods study could be conducted, but focusing on a broader and more established field than WAS.

It would be interesting to conduct a study with the objective of uncovering humanities research methodologies and recommending publisher infrastructure developments to accommodate the publication and open sharing of such methodologies. Such research would emphasise the importance of transparency and sharing of research methods and instruments, rather than the data itself.

Another useful study could be conducted regarding the legal restrictions (including privacy and third-party ownership), and the implications of using, managing, and sharing data in the humanities.

5.6 Study limitations

A notable limitation is that the current study is not generalisable due to its exploratory nature and small sample sizes. Second, there is an evident lack of previous research studies on this topic across this particular field of study in the Digital Humanities. Third, there is a lack of extant literature detailing publishers' activities regarding RDM and data sharing, and a lack of literature calling upon publishers to broaden the support they currently offer to Digital Humanities researchers. Fourth, self-selection bias was a limitation, as the sample did not adequately represent the population. Important to note, however, is that although the questionnaire respondents were few, many of their responses were in line with the literature, thus showing the questionnaire results to be reliable. Finally, there was lack of responses from researchers in South Africa, or Africa as a whole, which means drawing conclusions about WAS activities in this region was not possible.

5.7 Conclusion

Although this study's findings were mostly in line with the literature, not all the challenges evident in the literature are evident in the current research. Overall, the study found that WAS researchers tend to manage their data well. The main gaps in their current practices concern data sharing in formal repositories.

Given the various challenges that WAS researchers face when it comes to managing, and specifically sharing data, an approach different to those taken in the sciences is necessary if WAS research – and indeed research in the broader humanities and Digital Humanities – is to benefit from open scholarship. This will require considerable effort and investment on the part of publishers, as well as collaboration with other stakeholders in the research ecosystem. Without taking a different approach to making WAS research data as open as possible, data management and sharing support will remain patchy at best for researchers in this field, and will remain a lesser priority, halting development or progress in this endeavour. The world today relies on the internet to function: the archiving of information captured on the world wide web and making it accessible will greatly contribute to open research. Even if the data itself cannot be shared outright, an opportunity will be missed should there continue to be limited infrastructure supporting WAS researchers that prevents their research from being opened for validation and re-use at all possible stages in the research lifecycle.

References

ADHO. n.d. Announcing Liaison Between ADHO and The Research Data Alliance. Available: https://adho.org/announcements/2015/announcing-liaison-between-adho-and-research-data-alliance [2020, May 10].

Almas, B. 2017. Perseids: Experimenting with Infrastructure for Creating and Sharing Research Data in the Digital Humanities. *Data Science Journal*. 16(19): 1–17. DOI: 10.5334/dsj-2017-019.

Angen, M.J. 2000. Pearls, Pith, and Provocation: Evaluating Interpretive Inquiry: Reviewing the Validity Debate and Opening the Dialogue. *Qualitative Health Research*. 10(3): 378-395.

Association of College & Research Libraries. 2014. Top Trends in Academic Libraries: A Review of the Trends and Issues Affecting Academic Libraries in Higher Education. Available: https://crln.acrl.org/index.php/crlnews/article/view/9137/10062 [2020, October 17].

Ayris, P. 2017. Challenges and Opportunities for Research Data Management in the Arts, Humanities and Social Sciences: A Practitioner's Viewpoint. In LEARN Toolkit of Best Practice for Research Data Management. LEARN Project. 47–58. DOI: 10.14324/000.learn.00.

Babbie, E. 2016. *The Practice of Social Research*. 14th ed. Boston: Cengage Learning.

Banks, G.C., Field, J.G., Oswald, F.L., O'Boyle, E.H., Landis, R.S., Rupp, D.E. & Rogelberg, S.G. 2018. Answers to 18 Questions About Open Science Practices. *Journal of Business and Psychology*. DOI: 10.1007/s10869-018-9547-8.

Beagrie, N. 2019. *What to Keep: A Jisc Research Data Study.* Available: https://repository.jisc.ac.uk/7262/ [2019, March 13].

Bhattacherjee, A. 2012. *Social Science Research: Principles, Methods, and Practices*. Tampa: University of South Florida.

BioMed Central. n.d. Registered Reports. Available: https://www.biomedcentral.com/p/registered-reports [2021, January 06].

Borgman, C. 2009. The Digital Future is Now: A Call to Action for the Humanities. *Digital Humanities Quarterly*. 3(4). Available: http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html [2019, November 25].

Borgman, C. 2012. The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology*. 63(6): 1059-1078. DOI: 10.1002/asi.22634.

Bowen, G.A. 2009. Supporting a Grounded Theory with an Audit Trail: An Illustration. *International Journal of Social Research Methodology.* DOI: 10.1080/13645570802156196.

Brügger, N. & Laursen, D. 2019. Introduction: Digital Humanities, the Web, and National Web Domains. In *The Historical Web and Digital Humanities: The Case of National Web Domains*. N. Brügger & D. Laursen, Eds. London: Routledge. 1–9.

Brügger, N. & Laursen D. Eds. 2019. *The Historical Web and Digital Humanities*. Oxford: Routledge.

Brügger, N., Laursen, D., & Nielsen J. 2019. Establishing a Corpus of the Archived Web: The Case of the Danish Web from 2005 to 2015. In *The Historical Web and Digital Humanities: The Case of National Web Domains*. N. Brügger & D. Laursen, Eds. London: Routledge. 124–142.

Buddenbohm, S., Cretin, N., Dijk, E., Gaiffe, B., de Jong, M., Minel, J.L., Le Tellier-Becquart, N. 2016. State of the Art Report on Open Access Publishing of Research Data in the Humanities. DARIAH. Paris. Available: https://halshs.archives-ouvertes.fr/halshs-01357208v3/document [2021, February 10].

Carcary, M. 2009. The Research Audit Trail – Enhancing Trustworthiness in Qualitative Inquiry. *The Electronic Journal of Business Research Methods*. 7(1):11-24. Available: https://academic-publishing.org/index.php/ejbrm/article/view/1239 [2021, January 03].

Centre for Open Science [COS]. n.d.a. Open Science Badges. Available: https://www.cos.io/initiatives/badges [2020, October 17].

Centre for Open Science [COS]. n.d.b. Registered Reports: Peer Review Before Results are Known to Align Scientific Values and Practices. Available: https://www.cos.io/initiatives/registered-reports [2021, January 06].

Centre for Open Science [COS]. n.d.c. TOP Guidelines. Available: https://www.cos.io/initiatives/top-guidelines [2020, September 28].

Christian, T., Gooch, A., Vision, T. & Hull, E. 2020. Journal data policies: Exploring how the understanding of editors and authors corresponds to the policies themselves. *PLoS One*. DOI: 10.1371/journal.pone.0230281.

Clarivate Analytics. 2020. Web of Science Database. Available: https://apps.webofknowledge.com/WOS_GeneralSearch_input.do?product=WOS&search_mode=GeneralSearch&SID=D3GYjye7YTNSlHZ4i78&preferencesSaved= [2021, February 10].

CLOCKSS. n.d. Why CLOCKSS? Available: https://clockss.org/about/ [2021, January 06].

Cohen, P. 2010. Digital Keys for Unlocking the Humanities' Riches. *The New York Times*. 16 November 2010. Available: https://www.nytimes.com/2010/11/17/arts/17digital.html [2019, September 4].

Consortium of European Social Science Data Archives [CESSDA]. 2020. CESSDA Data Management Expert Guide. Available: https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide [2021, January 01].

Creswell, J.W. & Plano Clark, V.L. 2011. Choosing a Mixed Methods Design. In *Designing and Conducting Mixed Methods Research*. 2nd ed. J.W. Creswell & V.L. Plano Clark, Eds. Los Angeles: Sage.

Creswell, J.W. 2018. *Research Design: Qualitative, Quantitative and Mixed Methods Approaches*. Los Angeles: Sage.

DaMaRo. n.d. What is DaMaRo? Available: http://damaro.oucs.ox.ac.uk [2020, May 10].

Digital Research Infrastructure for the Arts and Humanities [DARIAH]. n.d.a. DARIAH Research Data Management Working Group. Available: https://www.dariah.eu/activities/working-groups/research-data-management/ [2020, May 10].

Digital Research Infrastructure for the Arts and Humanities [DARIAH]. n.d.b. Open Data for Humanists, A Pragmatic Guide. Available: https://dh.tcd.ie/dh/wp-content/uploads/2018/12/Open-Data-for-Humanists-A-Practical-Guide.pdf [2019, March 12].

"Data". Glossary. Digital Curation Centre. n.d. Available: http://www.dcc.ac.uk/digital-curation/glossary#D [2020, January 22].

Data Documentation Initiative [DDI]. n.d. DDI Lifecycle. Available: https://ddialliance.org/Specification/DDI-Lifecycle/ [2020, April 30].

Digital Curation Centre [DCC]. n.d.a. The DCC Curation Lifecycle Model. Available: http://www.dcc.ac.uk/resources/curation-lifecycle-model [2019, July 22].

Digital Curation Centre [DCC]. n.d.b. DMPOnline. Available: https://dmponline.dcc.ac.uk/ [2020, September 28].

DMPTool. n.d. Available: https://dmptool.org/ [2020, September 28].

Dressel, W.F. 2017. Research Data Management Instruction for Digital Humanities. *Journal of eScience Librarianship*. 6(2): e1115. DOI: 10.7191/jeslib.2017.1115.

Drucker, J. 2011. Humanities Approaches to Graphical Display. *Digital Humanities Quarterly*. 5(1). Available: http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html [2019, August 15].

Elsevier. n.d. Research Data Guidelines. Available: https://www.elsevier.com/authors/author-services/research-data/data-guidelines [2019, February 2].

F1000 Research. n.d. Publish Your Registered Report on F1000 Research. Available: https://think.f1000research.com/registered-reports/ [2021, January 06].

Faniel, I. M. & Zimmerman, A. 2011. Beyond the Data Deluge: A Research Agenda for Large-Scale Data Sharing and Reuse. *International Journal of Digital Curation*. 6(1): 58-69. DOI: https://doi.org/10.2218/ijdc.v6i1.172

Fear, K. 2015. Building Outreach on Assessment: Researcher Compliance with Journal Policies for Data Sharing. *Bulletin of the Association for Information Science and Technology*. 41(6): 18–21. DOI: 10.1002/bult.2015.1720410609.

Federer, L.M., Belter, C.W., Joubert, D.J., livinski, A., Lu, Y., Snyders, L.N. & Thompson, H. 2018. Data sharing in PLOS ONE: An Analysis of Data Availability Statements. PLoS One. 13(5). DOI: 10.1371/journal.pone.0194768

Fereday, J. & Muir-Cochrane, E. 2006. Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods*. 5(1): 80-92. DOI: 10.1177/160940690600500107.

Figshare. n.d. Figshare as an All-in-One Institutional Repository: Loughborough University's DSpace Migration, Reconfiguration of the Data Repository and Reintegration with Symplectic Elements. Available: https://knowledge.figshare.com/case-studies/figshare-as-an-all-in-one-institutional-repository-loughborough-universitys-dspace-migration-reconfiguration-of-the-data-repository-and-reintegration-with-symplectic-elements [2021, February 08].

Flanders, J. & Muñoz, T. n.d. An Introduction to Humanities Data Curation. In *DH Curation Guide: A Community Resource Guide to Data Curation in the Digital Humanities*. D.W. Anderson, K. Fenlon, M. Levine, C.M. Sperberg-McQueen, A. Babeu, J. Flanders, T. Muñoz, D. Dubin, J. Jett, & C.L. Palmer, Eds. Available: https://guide.dhcuration.org/contents/intro/ [2019, November 28].

Given, L.M. & Willson, R. 2018. Information Technology and the Humanities Scholar: Documenting Digital Research Practices. *Journal of the Association for Information Science and Technology*. 69(6): 807–819.

Goggin, G. & McLellan M. 2017. Introduction: Global Coordinates of Internet Histories. In *The Routledge Companion to Global Internet Histories*. H. Goggin & M. McLellan, Eds. London: Routledge. 1–19.

Gómez, N., Méndez, E., & Hernández-Pérez, T. 2016. Social Sciences and Humanities Research Data and Metadata: A Perspective from Thematic Data Repositories. *El Profesional de la Información*. 25(4): 545-555. DOI: 10.3145/epi.2016.jul.04.

Harrison, H., Birks, M., Franklin, R. & Mills, J. 2017. Case Study Research: Foundations and Methodological Orientations. *Forum: Qualitative Social Research*. *18(1): np*. Available: http://nbn-resolving.de/urn:nbn:de:0114-fqs1701195 [2019, August 31].

Henry, G. 2014. Data Curation for the Humanities: Perspectives from Rice University. In *Research Data Management: Practical Strategies for Information Professionals*. J. M. Ray, Ed. West Lafayette: Purdue University. 347–374. Available: https://www.jstor.org/stable/j.ctt6wq34t.20 [2020, January 13].

Herndon, J. & O'Reilly, R. 2016. Data Sharing Policies in Social Sciences Academic Journals: Evolving Expectations of Data Sharing as a Form of Scholarly Communication. In *Databrarianship: The Academic Data Librarian in Theory and Practice*. L. Kellam & K. Thompson, Eds. Chicago: American Library Association. 219–242.

Hockx-Yu, H., Laursen, D. & Gomes, D. 2019. The Curious Case of Archiving .eu. In *The Historical Web and Digital Humanities: The Case of National Web Domains*. N. Brügger & D. Laursen, Eds. London: Routledge. 64–72.

Hrynaszkiewicz, I. 2020. Accelerate Progress with a Journal Research Data Policy Framework. Available: https://www.stm-researchdata.org/wp-content/uploads/2020/02/Journal-policy-frameworks-STM-webinar-2020-2.pdf [2020, October 17].

Joint Information Systems Committee [Jisc]. n.d.a. RDM Checklist. Available: https://rdmtoolkit.jisc.ac.uk/plan-and-design/rdm-checklist/ [2020, October 12].

Joint Information Systems Committee [Jisc]. n.d.b. Research Data Lifecycle. Available: https://rdmtoolkit.jisc.ac.uk/research-data-lifecycle/ [2020, April 30].

Joint Information Systems Committee [Jisc]. n.d.c. Research Data Management Toolkit. Available: https://rdmtoolkit.jisc.ac.uk [2020, April 30].

Jones, L., Grant, R., & Hrynaszkiewicz. 2019. Implementing Publisher Policies that Inform, Support and Encourage Authors to Share Data: Two Case Studies. *Insights*. 32: 1-11. DOI: 10.1629/uksg.463

Kivunja, C. & Kuyini, A.B. 2017. Understanding and Applying Research Paradigms in Educational Contexts. *International Journal of Higher Education*. 6(5):26-41. DOI:10.5430/ijhe.v6n5p26.

Klein, L.F. & Gold, M.K. Eds. 2016. Digital Humanities: The Expanded Field. In *Debates in the Digital Humanities 2016*. Minneapolis: University of Minnesota Press.

Koopman, M. M. & De Jager, K. 2016. Archiving South African Digital Research Data: How Ready Are We? *South African Journal of Science*. 112(7): 2015-316. DOI:10.17159/sajs.2016/20150316.

Kvalheim, V. & Kvamme, T. 2014. Policies for Sharing Research Data in Social Sciences and Humanities: A Survey About Research Funders' Data Policies. International Federation of Data Organizations for Social Science. Available: https://www.cessda.eu/content/download/573/5371/file/ifdo_survey_report_2014.pdf [2019, March 29].

LOCKSS. n.d. What is LOCKSS? Available: https://www.lockss.org/about/what-lockss [2021, January 06].

MANTRA: Research Data Management Training. University of Edinburgh. 2017. Available: https://mantra.edina.ac.uk/ [2019, July 22].

Marr, B. 2018. How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. *Forbes*. 21 May. Available: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#63fcfbe860ba [2019, September 25].

Meyer, C.B. 2001. A Case in Case Study Methodology. *Field Methods*. 13(4): 329-352. DOI: 10.1177/1525822X0101300402.

Milligan, I. & Smyth, T.J. 2019. Studying the Web in the Shadow of Uncle Sam: The Case of the .ca Domain. In *The Historical Web and Digital Humanities: The Case of National Web Domains*. N. Brügger & D. Laursen, Eds. London: Routledge. 45–63.

Morse, J.M. 2012. Semistructured Interviews. In *The SAGE Handbook of Interview Research: The Complexity of the Craft*. J.F. Gubrium, J.A. Holstein, A.B. Marvasti & K.D. McKinney, Eds. 2nd ed. Washington: Sage. DOI: 10.4135/9781452218403.n13.

Munoz, T. & Renear, A. 2011. Issues in Humanities Data Curation. Available: http://cirss.ischool.illinois.edu/paloalto/whitepaper/premeeting/ [2019, November 29].

Naughton, L. & Kernohan, D. 2016. Making Sense of Journal Research Data Policies. *Insights*. 29(1): 84–89. DOI: 10.1629/uksg.284.

Ng'eno, E. & Mutula, S. 2018. Research Data Management (RDM) in Agricultural Research Institutes: A Literature Review. *Inkanyiso: Journal of Humanities and Social Sciences*. 10(1):28-50.

Organisation for Economic Co-Operation and Development [OECD]. 2007. OECD Principles and Guidelines for Access to Research Data from Public Funding. Available: http://www.oecd.org/sti/inno/38500813.pdf [2021, February 08].

Owens, T. 2011. Defining Data for Humanists: Text, Artifact, Information, or Evidence? *Journal of Digital Humanities*. 1(1). Available: http://journalofdigitalhumanities.org/1-1/defining-data-for-humanists-by-trevor-owens/ [2019, November 29].

Piwowar, H.A. 2011. Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. *PLoS ONE*. 6(7): 1-13. DOI: 10.1371/journal.pone.0018657.

Poole, A. 2013. Now is the Future Now? The Urgency of Digital Curation in the Digital Humanities. *Digital Humanities Quarterly*. 7(2). Available: http://www.digitalhumanities.org/dhq/vol/7/2/000163/000163.html [2019, November 29].

Poole, A.H. & Garwood, D.A. 2019. Digging into Data Management in Public-Funded, International Research in Digital Humanities. *Journal of the Association for Information Science and Technology*. DOI: 10.1002/asi.24213.

President's Council of Advisors on Science and Technology. 2007. Leadership Under Challenge: Information Technology R&D in a Competitive World – An Assessment of the Federal Networking and Information Technology R&D Program. Available at: https://www.nsf.gov/geo/geo-data-policies/pcast-nit-final.pdf [2019, September 25].

Primary Research Group. 2018. International Survey of Research University Faculty: Data Management and Archiving Needs. Report. Primary Research Group, Inc. Available: https://www.researchandmarkets.com/reports/4462900/international-survey-of-research-university [2021, February 10].

R Foundation. n.d. The R Project for Statistical Computing. Available: https://www.r-project.org/ [2021, 02 March].

Rendix, M. & Laursen, D. 2014. Digital Humanities: Now and Beyond. *MedieKultur: Journal of Media and Communication Research*. 57:1–3. Available: https://tidsskrift.dk/mediekultur/article/view/18610/17434 [2019, November 25].

Research Data Alliance [RDA]. n.d.a. About RDA. Available: https://www.rd-alliance.org/about-rda [2020, May 10].

Research Data Alliance [RDA]. n.d.b. Data Policy Standardisation and Implementation Interest Group. Available: https://rd-alliance.org/groups/data-policy-standardisation-and-implementation-ig [2019, July 22].

Research Data Management Librarian Academy [RDMLA]. 2019. Website homepage. Available: https://rdmla.github.io [2020, February 20].

Rousi, A. M. & Laakso, M. 2020. Journal Research Data Sharing Policies: A Study of Highly-Cited Journals in Neuroscience, Physics, and Operations Research. *Scientometrics*. 124:131-152. DOI: 10.1007/s11192-020-03467-9.

Schaffner, J. & Erway, R. 2014. Does Every Research Library Need a Digital Humanities Center? (Research report). Dublin, Ohio, United States of America: Online Computer Library Center (OCLC). Available: https://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-digital-humanities-center-2014.pdf [2019, April 1].

Schöch, C. 2013. Big? Smart? Clean? Messy? Data in the Humanities. *Journal of Digital Humanities*. 2(3). Available: http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/ [2019, November 29].

Schöpfel, J. & Prost, H. 2016. Research Data Management in Social Sciences and Humanities: A Survey at the University of Lille (France). *LIBREAS: Library Ideas*. 29: 98–112. Available: https://hal.univ-lille.fr/hal-01395816/document [2021, February 10].

Schroeder, R. & Brügger, N. 2017. Introduction: The Web as History. In *The Web as History: Using Web Archives to Understand the Past and the Presen*t. N. Brügger & R. Schroeder, Eds. London: UCL Press. 1–19. Available: https://mediarep.org/bitstream/handle/doc/4435/Bruegger_Schroeder_2017_The_Web_as_History_.pdf?sequence=3 [2020, January 13].

Shenton, A.K. 2004. Strategies for Ensuring Trustworthiness in Qualitative Research Projects. *Education for Information*. DOI: 10.3233/EFI-2004-22201.

Shipman, J.P. & Tang, R. 2019. The collaborative creation of a Research Data Management Librarian Academy (RDMLA). *Information Services & Use*. 39:243-247. DOI: 10.3233/ISU-190050.

Simms, S., Strong, M., Jones, S., & Riberio, M. 2016. The Future of Data Management Planning: Tools, Policies, and Players. *International Journal of Digital Curation*. 11(1): 208–217. DOI: 10.2218/ijdc.v11i1.413.

Springer Nature. n.d. Research Data Policy Types. Available: https://www.springernature.com/de/authors/research-data-policy/data-policy-types/12327096 [2019, February 2].

Sterba, S.K & Foster, E.M. 2008. Self-Selected Sample. In *Encyclopedia of Survey Research Methods*. P.J Lavrakas, Ed. Los Angeles: Sage.

Sturges, P., Bamkin, M., Anders, J.H.S., Hubbard, B., Hussain, A., & Heeley, M. 2015. Research Data Sharing: Developing a Stakeholder-Driven Model for Journal Policies. *Journal of the Association for Information Science and Technology*. 66(12): 2445-2455. DOI: 10.1002/asi.23336.

Swain, J. 2018. A Hybrid Approach to Thematic Analysis in Qualitative Research: Using a Practical Example. *Sage Research Methods: Cases in Sociology*. DOI: 10.4135/9781526435477. Postprint. Available: https://discovery.ucl.ac.uk/id/eprint/10042537/ [2020, October 09].

Taylor & Francis. n.d.a. Data Availability Statements. Available: https://authorservices.taylorandfrancis.com/data-sharing-policies/data-availability-statements/# [2021, February 08].

Taylor & Francis. n.d.b. Registered Reports at Taylor & Francis. Available: https://authorservices.taylorandfrancis.com/publishing-your-research/peer-review/registered-reports/ [2021, January 06].

Taylor & Francis. n.d.c. Understanding Our Data Sharing Policies. Available: https://authorservices.taylorandfrancis.com/understanding-our-data-sharing-policies/ [2019, February 2].

Teszelszky, K. 2019. The Historic Context of Web Archiving and the Web Archive: Reconstructing and Saving the Dutch National Web Using Historical Methods. In *The Historical Web and Digital Humanities: The Case of National Web Domains*. N. Brügger & D. Laursen, Eds. London: Routledge. 13–28.

The Open University. n.d. GDPR – How Does it Affect Research Data Management and Data Sharing? Available: http://www.open.ac.uk/library-research-support/sites/www.open.ac.uk.library-research-support/files/files/Guide_RDM_GDPR-HowDoesThisAffectRDM.pdf [2021, January 06].

UK Data Service. n.d. Research Data Lifecycle. Available: https://www.ukdataservice.ac.uk/manage-data/lifecycle.aspx [2020, April 30].

Ünal, Y., Chowdhury, G., Kurbanoğlu, S., Boustany, J., & Walton, G. 2019. Research Data Management and Data Sharing Behaviour of University Researchers. *Information Research*. 24(1): np. Available: http://www.informationr.net/ir/24-1/isic2018/isic1818.html [2019, September 23].

Vasileiou, K., Barnett, J., Thorpe, S. & Young, T. 2018. Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. *BMC Medical Research Methodology. 18. DOI:* 10.1186/s12874-018-0594-7.

Vasilevsky, N.A., Minnier, J., Haendel, M.A. & Champieux, R.E. 2017. Reproducible and Reusable Research: Are Journal Data Sharing Policies Meeting the Mark? *PeerJ*. 5. DOI: 10.7717/peerj.3208.

Vines, T. & Albert, A. 2020. The Effect of a Strong Data Archiving Policy on Journal Submissions (Part II). *The Scholarly Kitchen*. Available: https://scholarlykitchen.sspnet.org/2020/08/26/__trashed/ [2020, October 17].

Web ARChive Studies Network [WARCnet]. 2020. Working Groups. Available: https://cc.au.dk/en/warcnet/working-groups/ [2020, September 28].

Web of Science Group. n.d. Web of Science. Available: https://clarivate.com/webofsciencegroup/solutions/web-of-science/ [2019, September 26].

What is Humanities Research Data? Lafayette College. Available: https://dss.lafayette.edu/what-is-humanities-research-data/ [2018, November 18].

Wiley. n.d.a. Publish a Registered Report for an Early Peer Review of Your Proposed Research. Available: https://authorservices.wiley.com/author-resources/Journal-Authors/submission-peer-review/registered-reports.html [2021, January 06].

Wiley. n.d.b. Wiley's Data Sharing Policies. Available: https://authorservices.wiley.com/author-resources/Journal-Authors/open-access/data-sharing-citation/data-sharing-policy.html [2019, February 2].

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N. & Boiten, J.W. 2016. The Fair Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3: 160018. DOI: 10.1038/sdata.2016.18.

Wilms, L., Derven, C., O'Dwyer, L., Lingstadt, K. & Verbeke, D. 2019. Europe's Digital Humanities Landscape: A Study from LIBER's Digital Humanities & Digital Cultural Heritage Working Group. Report. Lille, France: Digital Humanities and Digital Cultural Heritage workshop at LIBER's 2018 Annual Conference.

Wilson, J., Patrick, M. & Rumsey, S. 2013. Introduction to Research Data Management – Case Studies (Oxford). DOI: 10.5281/zenodo.28326.

Winters, J. 2019. Negotiating the Archives of UK Web Space. In *The Historical Web and Digital Humanities: The Case of National Web Domains*. N. Brügger & D. Laursen, Eds. London: Routledge. 75–88.

Wolford, B. n.d. What is GDPR, the EU's New Data Protection Law? GDPR.eu. Available: https://gdpr.eu/what-is-gdpr/ [2021, January 06].

Wyatt, S. 2019. DARIAH Annual Event 2019: Humanities Data [Video file]. Available: https://www.youtube.com/watch?v=WiKwdienmjo [2019, August 15].

Zuckerman, H. & Ehrenberg, R.G. 2009. Recent Trends in Funding for the Academic Humanities and Their Implications. *Daedalus*. 138(1):124-146. DOI: 10.1162/daed.2009.138.1.124.

Appendix A (R1 interview questions)

1. Do you think the term 'data' can be used for research in the humanities as well as the sciences, and if so, can you provide some examples of humanities data?

2. Do you think that researchers in the humanities need to be supported differently to researchers in the sciences in terms of data management and sharing, and if so, how?

3. To what degree does your company currently support researchers during the planning and design, data collection, collaboration and data analysis stages of research (as reflected in the research data lifecycle)?

4. To what extent is your company involved in supporting researchers during the storage, archiving, and preservation stage of research (as reflected in the research data lifecycle)?

5. In what ways does it support researchers during the stages of publishing and sharing their data (as reflected in the research data lifecycle)?

6. To what extent does the company support researchers in discovering, re-using, and citing data?

7. Are there any stages in the research data lifecycle at which your company currently offers different support to humanities researchers than it does to researchers in other disciplines? If so, what are these stages, and how does the support differ?

8. At which stages of the research data lifecycle do you think publishers have the biggest responsibility or role to play, and why?

9. At which stages of the research data lifecycle do you think publishers do not have a responsibility or a role to play, and why?

10. In the company's engagement with data management and sharing, what challenges has it come across in planning and implementing initiatives, both generally and specifically for the humanities?

11. Are publishers doing any work to support researchers' data management and sharing practices in the field of the Digital Humanities specifically?

12. To what extent does your company promote data management and sharing internally to colleagues who either manage journals or have frequent contact with researchers and editors?

13. Does your company collaborate with other organisations or publishers in working towards a common objective to share data and support researchers – both more generally and with regard to the humanities specifically?

14. Bearing in mind the stages in the research data lifecycle, in what ways do you hope that humanities researchers will be supported by publishers in ten or twenty years from now?

15. Are there any final comments you would like to make, or any information you would like to share that hasn't been covered so far in this interview?

Appendix B (R2 interview questions)

1. As someone who specialises in the development of a portfolio of humanities, media, and arts journals in which a journal focusing on internet studies and web archives is situated, could you give a brief overview of what the field of Web Archive Studies entails?

2. Do you think the term 'data' applies to the kinds of research materials that Web Archive Studies researchers use?

3. How important or necessary do you think sound data management and data sharing is for this subject area?

4. At which stages of the research data lifecycle do you think publishers have the biggest responsibility or role to play, and why?

5. At which stages of the research data lifecycle do you think publishers do not have a responsibility or a role to play, and why?

6. At which stages of the research data lifecycle is the journal currently supporting Web Archiving Studies researchers, and how?

7. Your company offers a suite of standardised data sharing policies. Could you provide some detail on the standardised policy currently adopted by the journal in your portfolio that focuses on internet studies and web archives, including when it was adopted, and why?

8. Are journal authors encouraged to share their data at the point of article submission (during article file upload and acceptance of the publisher's terms and conditions)?

9. If an author makes it clear upon submission that there is a data set associated with their article, but a Data Availability Statement is not included in the submission, is there a process in place whereby the publisher will solicit this statement from the author?

10. There is a journal in your portfolio that is situated in the field of Web Archive Studies. Since it was launched, do you know how many authors have shared their data, and how many authors have included a Data Availability Statement in their published articles (if the numbers are different)?

11. For the researchers who do share data in the journal, can you share with me the sense you have of the following:

    a. Do the researchers collect their own data, or do they tend to re-use data that has been collected by others?
    b. What types of data do the researchers generate or use, and in which formats (i.e. voice recordings as .mp3 files, typed out transcripts in Word, numerical data in spreadsheets, graphical images as .tiff files, PDF graphs, digital photographs as JPEGs, etc.)?
    c. Where do the researchers tend to store their data, and does your company provide assistance with this?

12. Are studies by Web Archive Studies researchers generally funded by organisations that mandate data sharing?

13. Do any of the other journals in your broader field of media, communication, and cultural studies portfolio have a stricter data sharing policy, or engage with data sharing in a way that other journals do not typically do?

14. Apart from the data sharing policy, does the journal support and encourage authors in managing and/or sharing their data via other initiatives or methods such as author workshops, information packs, or webinars?

15. Bearing in mind the different stages in the research lifecycle, are you aware of any specific challenges that Web Archive Studies researchers might commonly face in terms of data management and data sharing, such as privacy or intellectual property?

16. Have you received any training or information internally about the importance and necessity of data management and data sharing, and the role that your company currently plays in the wider field of open scholarship?

17. In what ways do you hope that Web Archive Studies researchers will be supported by publishers in ten or twenty years from now, with regard to data management and sharing?

18. Are there any final comments you would like to make, or any information you would like to share that hasn't been covered so far in this interview?

Appendix C (R3 interview questions)

1. In your understanding, what does the field of Web Archive Studies entail, and how does it fit into the broader field of the Digital Humanities?

2. What do you understand by the term 'data', and does the term apply to the kinds of research materials that Web Archive Studies researchers use?

3. How important or necessary do you think sound data management and data sharing is for this subject area?

The following questions relate to your own research data management and sharing practices during each of the six stages in the research data lifecycle:

4. Stage 1: Planning and designing

    a) In your research, do you usually create a data management plan (DMP) at the beginning of a study? If so, is this due to a funder or institutional policy that mandates such plans?
    b) What challenges do you face during the data planning stage?

5. Stage 2: Data collection and capture

    a) What types of data do you usually generate or use to develop a corpus, and in which data formats (i.e. html code as a .doc file, voice recordings as .mp3 files, typed out transcripts in Word, numerical data in spreadsheets, graphical images as .tiff files, PDF graphs, digital photographs as JPEGs, etc.)?
    b) What sizes of data do you work with (i.e. how is the collected data quantifiable in terms of MB, GB, TB)?
    c) What kind of file naming conventions do you use when gathering your data?
    d) How and where do you store your data during the data collection stage?
    e) Do you assign metadata to the collected data, and if so what kind of metadata do you include (e.g. title, author, subject, keywords, publisher, description)?
    f) What challenges do you face during the data collection and capture stage?

6. Stage 3: Collaboration and analysis

    a) Do you typically conduct your research alone or in a team of researchers? If with a team of researchers, what is the size of the group?
    b) Do you keep a record of your data analysis methods, and if so, how?
    c) What challenges do you face during the collaboration and analysis stage?

7. Stage 4: Data management, storage, and preservation

    a) How do you back-up your data during and after collection?
    b) Are there any security regulations that affect how you store your data?
    c) What other challenges do you face during the data management, storage, and preservation stage?

8. Stage 5: Data sharing and publication

a) Do you share your research data via more informal modes such as email, Google docs, or file transfer services? If so, which ones?
b) Do you share your research data more formally in repositories that assign persistent identifiers or DOIs to the data? If so, which ones?
c) To what extent does intellectual property and licensing affect your ability to share research data?
d) What other challenges do you face during the data sharing and publication stage?

9. Stage 6: Discovery, re-use, and citation

    a) Do you typically generate your own data, or do you use data that already exists and was generated by parties outside of your research team?
    b) Do you use your own pre-existing data for future research studies?
    c) If you use pre-existing data (either generated by other parties or by yourself), do you typically require permission to do this, or is the data licenced for re-use?
    d) If you use pre-existing data, what challenges do you face in either finding, accessing, re-using, or citing the data?

10. At what stages of the research data lifecycle do you think publishers specifically could provide additional support to alleviate some of the challenges you've mentioned? What kind of support would this involve?

11. At which stages of the research data lifecycle do you think publishers do not have a role to play in supporting researchers, and why?

12. Are you aware if studies in Web Archive Studies are generally funded by organisations that mandate data sharing?

13. Are you involved in any initiatives, groups, or organisations that work towards sound data management and sharing in the field of Web Archive Studies specifically?

14. As the editor of a journal focusing on internet studies and web archiving, do you actively encourage authors to share the data associated with their published research articles?

15. As the editor of the journal, have you received any support from the publisher in encouraging and promoting authors to share data?

16. The publisher of the journal focusing on internet studies and web archiving offers a suite of standardized data sharing policies. The journal currently uses the most basic of these policies, which encourages but does not require authors to share their data. Do you think this policy is adequate for the journal? Do you think the journal could benefit from a stricter policy?

17. Are you aware of any training, information packs, webinars, or other resources provided by publishers more generally to assist researchers with data management or sharing?

18. In what ways do you hope that Web Archive Studies researchers will be supported by publishers in ten or twenty years from now, with regard to research data management and sharing?

19. Are there any final comments you would like to make, or any information you would like to share that hasn't been covered so far in this interview?

Appendix D (Questionnaire)

Introduction

Thank you for agreeing to participate in this study.

This questionnaire is intended only for respondents who are currently active researchers in the field of Web Archive Studies. Web Archive Studies is an interdisciplinary field that explores the historical, social and cultural implications of information published on the internet.

What is the nature of my research study?
My research study investigates Research Data Management (RDM) and data sharing in the digital humanities. A case study will be developed which focuses specifically on RDM and data sharing in the field of Web Archive Studies. The overall objective of the study is to recommend ways that publishers can better support researchers in this field in managing and sharing their data.

What does your participation involve?
Participation in the study will involve you answering a series of questions about your research data management (RDM) and data sharing practices.

- *Risks:* There are no risks involved in participation.
- *Benefits:* A possible benefit for participants is that recommendations made to publishers are implemented, thus providing more data management and sharing support to Web Archive Studies researchers.
- *Costs:* Costs for the participant include the ability to connect to the questionnaire via SurveyMonkey, and the time spent completing it.

How will your responses be used?
Responses will be used to understand Web Archive Studies researchers' current research data management and sharing practices. Responses will inform recommendations as to how publishers can better support researchers in this field.

Responses will be used for education and research purposes only. Respondents will remain anonymous, and any personal information will remain strictly confidential.

Questions

Part 1: Researcher Information

The following information is needed only to confirm that you are an active researcher working in the field of Web Archive Studies. Even if your research isn't typically labelled as Web Archive Studies, it will likely fall into this field if your work concerns the historical, cultural, or social implications of information published on the internet. Your answers will remain strictly confidential.

Question 1
Please state in the text box below the academic department or research centre with which you are affiliated, as well as the nature of your research as related to Web Archive Studies:

[Text box for response]

Question 2

Have you published an article in an academic journal, conference proceedings in a compilation, or a book chapter in the last 18 months?

□ Yes
□ No

Part 2: Attitudes Toward Research Data Management in the Field of Web Archiving Studies

Research data refers to a broad range of research materials that are collected and used as evidence on which to base a finding. Some examples of research data include HTML code, numerical data, voice recordings, digital images, text documents, field notes, or graphs.

Question 3
How important do you think research data management is for research studies in Web Archive Studies? Please select one answer:

□ Very important
□ I'm unsure
□ Not important at all

Question 4
How important is it for Web Archive Studies researchers to share their data with others? Please select one answer:

□ Very important
□ I'm unsure
□ Not important at all

Part 3: Types and Sizes of Data

The following questions relate to the kinds of data used in your research, as well as the size or amounts of data you use.

Question 5
Although researchers in all disciplines will use data in their studies, not all researchers use the term 'data'. For example, some humanities researchers might prefer the term 'research materials'. How would you typically refer to data in the field of Web Archiving Studies? Please select all that apply:

□ Data
□ Research materials
□ Other (please specify)

If 'Other' option chosen, provide text box:

[Text box for response]

Question 6
What kinds of data do you collect and use for your research? Please select all that apply:

□ Text files
□ Images

□ Audio files
□ Numerical data (e.g. statistics)
□ Graphical images
□ HTML code
□ Geospatial data
□ Archival metadata
□ Crawl logs
□ Publications (e.g. journal articles or books)
□ Field notes
□ Other (please specify)

If 'Other' option chosen, provide text box:

[Text box for response]

Question 7
What sizes or amounts of data do you usually collect and use? Please select all that apply:

□ KB (kilobytes)
□ MB (megabytes)
□ GB (gigabytes)
□ TB (terabytes)
□ Other (please specify)

If 'Other' option chosen, provide text box:

[Text box for response]

Question 8
What challenges have you experienced with regard to the kind of data that you use, as well as the size? Please select all that apply:

□ I have experienced no challenges.
□ The raw data was too large in size for my computer to process.
□ The raw data was duplicated in certain instances.
□ The original data set was too small to base a study on.
□ Other (please specify)

If 'Other' option chosen, provide text box:

[Text box for response]

Part 4: Size of Researcher Team

Question 9
Do you conduct your research alone or with a research team? Please select one answer:

□ Alone
□ With a research team
□ It depends

If ‘With a research team’ or ‘It depends’ chosen, Questions 10-12 should be asked.
If ‘Alone’ chosen, jump to Question 13.

Question 10
What are the sizes of the research teams you are usually involved with? Please select all that apply:

□ 1-10 research team members
□ 10-20 research team members
□ 20-30 research team members
□ 30-50 research team members
□ Over 50 research team members

Question 11
How do you collaborate with research team members? Please select all that apply:

□ Through email
□ Through collaborative tools (e.g. Google Docs, Slack, WhatsApp, OneDrive, etc)
□ Through digital meetings (e.g. using Zoom, Microsoft Teams, Skype, etc).
□ In person at meetings
□ Other (please specify)

If ‘Other’ option chosen, provide text box:

[Text box for response]

Question 12
What are your challenges involved with working with multiple research team members? Please select all that apply:

□ Coordinating multiple team members.
□ Delegating work fairly.
□ Working across time zones
□ The systems used to collaborate are not adequate for our purposes
□ The size of the team
□ Other (please specify)

If ‘Other’ option chosen, provide text box:

[Text box for response]

Part 5: Data Planning

The following questions relate to the methods and processes used when planning for a research project and outlining the data needed.

Question 13
Have you ever had to create a Data Management Plan for your research?

□ Yes
□ No

If ‘Yes’ chosen, Questions 14-15 should be asked.
If ‘No’ chosen, jump to Question 16.

Question 14
Why did you create a Data Management Plan? Please select all that apply:

□ Institutional requirement
□ Funder requirement
□ Researcher’s choice

Question 15
What specific challenges did you experience when creating your Data Management Plan? Please select all that apply:

□ Unfamiliarity with the purpose of a Data Management Plan.
□ Uncertainty about what to include in the Data Management Plan.
□ Issues with Data Management Planning tools and systems (e.g. DMPTool, DMP Online, etc).
□ Other (please specify)

If ‘Other’ option chosen, provide text box:

[Text box for response]

Question 16
How do you think a Data Management Plan might have helped you or your research team in past research studies? Please respond in the text box below:

[Text box for response]

PART 6: Data Collection and Analysis

The following questions concern the data collection and analysis stage of research.

Question 17
Have you or your research team members used secondary data (pre-existing data generated by a third party) in your research?

□ Yes
□ No

If ‘Yes’ chosen, Questions 18-20 should be asked.
If ‘No’ chosen, jump to Question 21.

Question 18
How did you find and access the data set? Please select all that apply:

□ I found the data in a data repository
□ The data was emailed to me
□ The data was shared with me via a file sharing service (e.g. Google Docs, WeTransfer, etc)
□ Other (please specify)

If ‘Other’ option chosen, provide text box:

[Text box for response]

Question 19
How did you attain permission to use the dataset? Please select all that apply:

□ The dataset was licensed for re-use (e.g. it was assigned a Creative Commons license allowing for re-use).
□ I obtained permission directly from the person or organisation who owns the data.
□ Other (please specify)

If ‘Other’ option chosen, provide text box:

[Text box for response]

Question 20
Did you cite the dataset in a publication (e.g. scholarly journal article, conference paper, book chapter, etc)? Please select one answer:

□ Yes
□ No

Question 21
Have you or your research team members generated new data for your research?

□ Yes
□ No

Question 22
Which of the following file naming conventions did you or your research team include when collecting and saving data? Please select all that apply:

□ Dates
□ Places
□ Numbers
□ Event
□ Comments
□ Initials
□ Other (please specify)

If ‘Other’ option chosen, provide text box:

[Text box for response]

Question 23
Have you or your research team used a pre-existing method, process, workflow, or model to analyse your data? Please select all that apply.

□ Method

□ Process
□ Workflow
□ Model
□ Other (please specify)

If 'Other' option chosen, provide text box:

[Text box for response]

Question 24
How did you keep a record of your data collection and analysis methods? Please select all that apply:

□ Maintained version control of all files
□ Kept a research diary or notebook
□ Kept photographic records
□ Kept audio recordings
□ Kept a time log
□ Other (please specify)

If 'Other' option chosen, provide text box:

[Text box for response]

Question 25
What challenges did you experience with the equipment that was available to you for data collection or analysis? Please select all that apply:

□ I experienced no challenges.
□ Computer hardware
□ Computer software
□ Internet speed
□ Lack of storage space
□ Other (please specify)

If 'Other' option chosen, provide text box:

[Text box for response]

Question 26
What other challenges have you experienced with regard to data collection and analysis? Please select all that apply:

□ I have experienced no challenges.
□ I didn't have enough time for this stage in the research.
□ I struggled to get permission to use third-party data.
□ Other (please specify)

If 'Other' option chosen, provide text box:

[Text box for response]

Part 7: Data Storage

The following questions concern the ways in which researchers store their data after it is generated.

Question 27
Where do you store your data? Please select all that apply:

□ Internal computer hard drive
□ External computer hard drive or flash drive
□ The cloud (e.g. Google Drive, DropBox, or iCloud)
□ Other (please specify)

If 'Other' option chosen, provide text box:

[Text box for response]

Question 28
How do you back up your data for long-term storage after you have collected it? Please select all that apply:

□ I do not back up my data
□ Internal computer hard drive
□ External computer hard drive or flash drive
□ The cloud (e.g. Google Drive, DropBox, or iCloud)
□ Printed documents
□ Other (please specify)

If 'Other' option chosen, provide text box:

[Text box for response]

Question 29
How much of your collected data is stored long-term? Please select one answer:

□ All of it
□ Some of it

Question 30
How long do you intend to keep your data once a research study is complete? Please select all that apply:

□ Less than 1 year
□ Between 1-2 years
□ Between 2-5 years
□ Between 5-10 years
□ Longer than 10 years

Question 31
What challenges do you experience with regard to storing your data? Please select all that apply:

□ I have experienced no challenges

□ Security challenges
□ Privacy challenges
□ Intellectual property challenges
□ Funding challenges
□ Storage space challenges
□ Other (please specify)

If 'Other' option chosen, provide text box:

[Text box for response]

Part 8: Data Sharing

The following questions concern the sharing and publication of research data.

Question 32
Have you shared your data in an institutional or subject repository?

□ Yes
□ No

If 'No' chosen, Question 33 should be asked.
If 'Yes' chosen, jump to Question 34.

Question 33
Which repository did you share your data in?

[Text box for response]

Question 34
In what other ways have you shared your research data with others? Please select all that apply:

□ Email
□ File sharing services (e.g. Google Drive, iCloud)
□ File transfer services (e.g. WeTransfer)
□ Social media (e.g. Facebook, Twitter, WhatsApp)
□ Other (please specify)

If 'Other' option chosen, provide text box:

[Text box for response]

Question 35
Was your data assigned a persistent identifier (such as a DOI) when you shared it?

□ Yes
□ No

Question 36
Metadata provides a description of research data, allowing research data to be searchable in a repository or database. What kind of information did you or your research team members include in the metadata describing your data? Please select all that apply:

□ Title
□ Author
□ Subject
□ Keywords
□ Publisher
□ Description
□ No metadata was generated

Question 37
What challenges did you experience with regard to sharing your data? Please select all that apply:

□ I have experienced no challenges
□ Security challenges
□ Privacy challenges
□ Intellectual property challenges
□ Funding challenges
□ Storage space challenges
□ Other (please specify)

If 'Other' option chosen, provide text box:

[Text box for response]

Part 9: Data Publication

Question 38
When publishing your research, has a publisher ever required that you submit a Data Availability Statement?

□ Yes
□ No

Question 39
When publishing, have you ever had to adhere to a data sharing policy that required you to share your data?

□ Yes
□ No

Question 40
Have you ever been required by a publisher to submit your Data Management Plan before acceptance/publication of your research?

□ Yes
□ No

Part 10: Data Re-Use

Question 41
How long after you have collected your data do you re-use the data? Please select all that apply:

□ Within 12 months of the data being collected.
□ Between 1-2 years after the data is collected.
□ Between 2-5 years after the data is collected.
□ Between 5-10 years after the data is collected.
□ After 10 years of the data is collected.
□ I don't re-use my data.

Question 42
Has your data been used by anyone else?

□ Yes
□ No
□ I don't know

Part 11: *Support for Researchers*

The following questions relate to Research Data Management and data sharing support for researchers, and where such support can be found.

Question 43
Have you ever received any training from academic journal editors or publishers to assist you with managing or sharing your data?

□ Yes
□ No

If 'Yes' chosen, Question 44 should be asked.
In 'No' chosen, jump to Question 45.

Question 44
What kind of training have you received from academic journal editors or publishers? Please select all that apply:

□ Workshop
□ Webinar
□ Training session
□ Other (please specify)

If 'Other' option chosen, provide text box:

[Text box for response]

Question 45
Have you ever been provided with any resources or guidelines by academic journal editors or publishers to assist you with managing or sharing your data?

□ Yes
□ No

If 'Yes' chosen, Question 46 should be answered.

If ‘No’ chosen, jump to Question 47.

Question 46
What kind of resources or guidelines have been provided to you by academic journal editors or publishers to assist you with managing or sharing your data? Please select all that apply:

□ General data sharing guidelines
□ Information on how to choose an appropriate repository for my research data
□ Other (please specify)

If ‘Other’ option chosen, provide text box:

[Text box for response]

Question 47
What additional support from academic journals and publishers would you find most useful as an author of Web Archiving Studies research? Please select all that apply:

□ I would like to receive Research Data Management training from a publisher.
□ I would like to receive data sharing training from a publisher.
□ I would like to receive information on which repositories I should use for my subject area.
□ I would like to receive written information on the benefits of Research Data Management and data sharing (e.g. a handbook, online toolkit resources).
□ I would like publishers to be more involved in Research Data Management and data sharing from an early stage in the research.
□ I would like to see publishers collaborating more with institutions and libraries to offer more support to researchers in managing and sharing their data.
□ I think that publishers should introduce stricter Data Sharing Policies to ensure that researchers in Web Archiving Studies are making their data available to others.
□ Other (please specify)

If ‘Other’ option chosen, provide text box:

[Text box for response]

Question 48
What support have you received from your institution to assist you with managing or sharing your data? Please select all that apply:

□ I have received training on Research Data Management
□ I have received training on data sharing
□ I have received training on how to complete a Data Management Plan
□ I have been made aware of repositories where I can share my data
□ I have received no training
□ Other (specify)

If ‘Other’ option chosen, provide text box:

[Text box for response]

Part 12: Final Comments

Question 49

Are there any final comments you would like to make, or any other challenges you would like to highlight in your experience of managing and sharing data? Please type in the textbox below:

[Text box for response]

Appendix E (Consent Form)

Title of research study:

Research Data Management and Sharing Practices in the Digital Humanities with a Focus on Publisher Support: A Case Study in the Field of Web Archive Studies

Nature of the research:

The research study investigates Research Data Management (RDM) and data sharing in the digital humanities. A case study will be developed which focuses specifically on RDM and data sharing in the field of Web Archive Studies.

The study seeks to understand:

- the current RDM and data sharing practices of Web Archive Studies researchers, and
- how publishers are currently engaging with RDM and data sharing, both generally and with regard to the digital humanities.

The overall objective of the study is to recommend ways that publishers could support digital humanities researchers in managing and sharing their data, specifically those in the field of Web Archive Studies.

The study involves three interviews (one with you as the participant), as well as a questionnaire that will be sent to Web Archive Studies researchers.

You are free to leave the study at any time.

Participant's involvement:

What does the participation involve? You will be interviewed using a semi-structured interview, allowing for follow up questions. You will be interviewed regarding researcher data management and sharing practices, as well as the data management and sharing support and engagement on the part of publishers, with a particular focus on Web Archive Studies researchers and publications. The interview questions have been sent to you previously. The interview will be conducted and, with permission, recorded virtually on a program called Microsoft Teams, and is estimated to last 60-90 minutes.

- *Risks:* There are no risks involved in participation.
- *Benefits:* Possible benefits for the participant involve relevant recommendations for publishers to improve their data management and sharing support for Web Archive Studies researchers. This would potentially allow for an improved author experience.
- *Costs:* Costs for the participant include the ability to connect to a virtual call via Microsoft Teams, and the time spent during the interview.

Name, signature and consent of participant:

- I agree to participate in this research project.
- I have read this consent form and the information it contains and had the opportunity to ask questions about them.
- I agree to my responses being used for education and research on condition that my privacy is respected, subject to the following:
    - I understand that I am under no obligation to take part in this project.
    - I understand I have the right to withdraw from this project at any stage.
    - I understand that this research might be published in a research journal. In the case of dissertation research, the document will be available to readers digitally in a university repository.
- I grant permission for the interview to be recorded:
  Yes □ No □
- I grant permission for my name and affiliation to be included in the research:
  Yes □ No □

Name of participant: _____

Signature of participant: _____

Date: _____

Name and signature of person who sought consent:

Name: _____

Signature: _____

Date: _____