Re-envisioning access for the digital preservation community: challenges, opportunities and recommendations

Thesis submitted in fulfilment of the degree of PhD in Digital Humanities

Leontien Talboom

UCL

2022

Declaration

I, Leontien Kate Talboom, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I can confirm that this has been indicted in the thesis.

Abstract

Digital material is not new and has been preserved for a couple of decades now. With a growing digital preservation community, and a growing number of practitioners identifying as doing something digital, there is an understanding that this material is here to stay. More and more institutions are publishing digital strategies or creating networks focusing on digital material. However, when looking at this in practice there seems to be a disconnect between what is being stated within these networks and strategies and what is being made accessible to the public.

This thesis will explore this disconnect by first understanding how the digital preservation community has been providing access to this material and how they are envisioning it in the future. This exploration surfaces both a) how digital material can no longer be seen as separate from the infrastructure that ensures its materiality and b) how the provision of access is not just a technological question, but also a social, legal and ethical one.

This thesis will also seek to explore the ways in which those who identify as digital preservation practitioners articulate their role and responsibilities. It will do so by drawing on relevant literature and gaining perspectives from practitioners and other relevant participants through in-depth interviews. Building from this exploration, this thesis will offer recommendations for how this practice can move forward in negotiating the provision of access to digital material in the online public space of the internet.

This research is part of a collaborative project with The National Archives, UK where a number of the ideas encountered during this work were explored in practice. Some of these results have helped shape the recommendations given in the final chapters of this thesis.

Impact Statement

This work contributes to the understanding of digital preservation practice. It starts by mapping out the current provision of access that digital preservation practitioners are providing, not only from a general point of view, but also by gaining perspectives from interviewed participants. A similar process is carried out in respect of the role and responsibilities of digital preservation practitioners. In this way it creates a better view of how access could be provisioned within the new space of the internet and how the use of technology can be optimised by working with more technical professionals, whilst still respecting the societal mission of the digital preservation practitioner. The perspective of data journalists has been introduced as a comparator and this introduction serves to bring to light an emphasis on the importance of spreading a societal mission in the online public space and on how Big Tech does not have to be followed to achieve this. This overview and discussion is not only useful for the digital preservation community itself, but also for wider humanities practice, as these individuals are preserving digital material for the future.

This research is a collaborative PhD between University College London and The National Archives (TNA) in the UK. The collaboration meant that there was room to experiment with a number of findings in practice. These outputs include a Machine Learning Club with Mark Bell at TNA, on which a book chapter has been published and a blog post discussing the concept of the designated community with David Underdown. In addition, experimentation has been carried out around providing access to archival material as data, with a focus on the UK Government Web Archive (UKGWA). This work has included a Data Study Group with the Alan Turing Institute, a data workshop with the Computational Archival Science group and an article on working with the UKGWA material. During the course of this PhD a Data Science and Humanities group was run in collaboration with the Alan Turing institute and the topics discussed by that group were heavily influenced by this work. Finally, a project that is currently running and which has benefited from the findings of this work is a computational access guide that is being written for the digital preservation community in collaboration with the Digital Preservation Coalition and is part of a Software Sustainability Institute fellowship.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisors, Jenny Bunn and Antonis Bikakis. Thank you so much for all your help, perseverance, and enthusiasm throughout my work. Thank you for keeping me on track and not letting me go off on too many random tangents. I would also like to thank Alec Mulinder, even though you might not have been able to finish my supervision, your excitement and willingness to help were a great start for my PhD journey.

Besides my supervisors I would like to thank everyone at The National Archives who got involved in my work, especially Mark Bell. Thank you so much for being so welcoming and helpful, especially when contacting participants and when I was stuck on ideas. The Information Studies department at UCL has also been of great help, especially my own PhD cohort and our graduate tutor Annemaree Lloyd.

Furthermore, I would like to thank all of my participants, not only for their enthusiasm and willingness but also for their persistent interest in my work. I would also like to thank my funding body LAHP, without them this PhD would have never happened. A special thanks to them for creating such an interesting PhD structure, as the collaborative nature of this PhD made it a very enjoyable experience.

Last but not least, I would like to thank Ben. Thank you so much for listening to my many ideas and rants throughout my PhD, it feels like you yourself have become an expert in this very niche topic of the provision of access by digital preservation practitioners.

1 Introduction

Digital material is nothing new and has been preserved for decades now. Early examples date back to the 1960s when the need for preservation arose from the fact that digital material was increasingly being re-used in research carried out in the social sciences (The Steinmetz Archive 1989; GESIS – Leibniz-Institut für Sozialwissenschaften 2019; ICPSR 2019). From the 1970s onwards other types of material started to be preserved including electronic text archives and other forms of historical data (Doorn and Tjalsma 2007). However, it was not until the 1990s when the first steps towards formalising digital preservation needs were made, with reports and articles being released on the concept and growing need for digital preservation (Garrett and Waters 1996; Hedstrom 1998; CCSDS 2002).

The release of these first reports led to more discussion and formalisation of digital preservation. The importance of this digital material started to be acknowledged, with models such as Open Archival Information System (OAIS) being widely adopted by institutions looking after digital material (van Essen 2019). At this point there was a strong emphasis on the preservation of this material, driven by a very real concern that it would be lost to technological obsolescence very quickly (Hedstrom 1998) and different approaches to preservation were established (Gollins 2009; Digital Preservation Coalition 2015).

The growth in digital material continued into the early 2000s, not just with institutions starting large scale digitisation projects, but also with the increasing number of digital records being created. As the discussions around digital preservation continued there was still a strong emphasis on the preservation of this material, but not necessarily on providing access to it. However, a shift can be observed from around the mid-2010s when larger institutions within the field started to take a more holistic approach towards this digital material.

Since these earlier years, a number of institutions have published digital strategies, outlining their ambitions for this type of material and not only focusing on its selection and preservation, but also on making it more accessible and available for use (The British Library 2017; The National Archives 2017b). During the same time period, networks focusing on this material such as the Dutch Digital Heritage Network (Netwerk Digitaal Erfgoed 2019) and the Computational Archival Science network (Goudarouli 2018) have been created and some Galleries, Libraries, Archives and Museums (GLAM) institutions have established labs to allow them to play around and discover the possibilities of digital material (Mahey et al. 2019).

Nowadays it seems that many institutions see the importance of this material and have ideas and strategic ambitions for providing access to it. However, when looking at what is actually happening in practice there seems to be a disconnect between ambition and reality. In many cases access to digital material is only provided through online catalogues, where this material can be difficult to navigate and be discovered in the first place (Gollins and Bayne 2015, 131–33). Sometimes this material is only viewable in physical reading rooms on predesigned online environments, and sometimes it is not accessible at all. Examples of this are the UK Web Archive (UKWA), which under Non-Print Legal Deposit Regulations from 2013 can only be viewed in the physical reading rooms (British Library 2021b), the Transport for London (TfL) Corporate Archives where only descriptions of digital material can be searched and consulted online (Transport for London 2021), and the Discovery catalogue from The National Archives (TNA) where although a substantial amount of items are viewable online, others are not; as an example, the Derbyshire collection is missing publicly open material that can neither be viewed online nor on site (The National Archives 2021b).

This leads to the main motivation for this thesis. In the last few years the importance of preserving and providing access to digital material has been understood and many projects, networks and strategies have discussed the possibilities for making this happen. However, this does not seem to coincide with what is currently happening in practice where access to such material is still only partial and difficult at best. This research explores how the digital preservation community is currently thinking about access and explore where this disconnect between what is being said and what is happening in practice is coming from.

The focus of this work is mainly on digital preservation practitioners looking after cultural heritage material and are therefore based at GLAM institutions. It should be acknowledged that other individuals may also identify as digital preservation practitioners, as is evident from the members list of the Digital Preservation Coalition (DPC) (Digital Preservation Coalition 2022a). The main questions around the difficulties of access are coming from the GLAM institutions, who in many cases have a public mandate to make this material accessible to a wide range of users.

It should also be noted that the research mainly focuses on activities in Europe, North America and Australia and with more emphasis in the Netherlands and the United Kingdom (UK). In short, this choice was made as these were feasible areas for the researcher to travel to and visit institutions. Furthermore, through the collaborative nature of the project it was possible to get in contact with a large range of institutions within the UK. The choice to

expand to the Netherlands was made as the researcher has a background in the Netherlands. Due to the pandemic the pool of researchers was extended outside of this range to include America and Australia, this is further detailed in Chapter 2.

1.1 Research Question

The main aim of this research is to understand the disconnect between what digital repositories are aiming to provide in terms of access, and what is being provided in practice, which leads to the following research question:

'How is digital preservation practice negotiating the provision of access in the digital environment?'

This thesis will be focusing on digital preservation practice, and therefore on digital preservation practitioners. The term digital preservation practitioner was chosen over archivist, as it is not only those who identify as archivists who preserve digital material. Across the humanities sector, including all GLAM institutions, but also within the fields of social and hard sciences, digital material is being preserved. Not all custodians of this material refer to themselves as archivists; many do not have a background within the archival field.

I personally identify as a digital preservation practitioner, having worked at the Archaeology Data Service (ADS) before starting this PhD work. I do not have a background in archives and records management, but instead in archaeology and computer science. Nonetheless I do feel a part of and able to actively contribute to the digital preservation practitioner community, a more detailed discussion of which can be found in Chapter 4.

What the provision of access means is very difficult to define and will be further explored during this research. As a starting point, the assumption is that at a bare minimum users should be able to access the material in some shape or form in an online public space (Appel et al. 2015, 16). This assumption would seem to be a commonly held one, since the digital strategies published by larger repositories mention online or remote access via the internet. An example of this is in the guidelines that TNA have published as part of their digital capacity building strategy called *'Plugged In, Powered Up'.* This strategy recommended that access should be provided ideally through an online interface (The National Archives 2019). The Canadian Centre for Architecture (CCA), who created SCOPE, which a browser-based access interface for digital material, also advocate for an online or remote access approach (Stewart and Breitweiser 2019).

The starting assumption is not however that the provision of access should imply open access. The open access movement aims to make research outputs openly accessible, which means that open access literature is digital, online, free of charge, and free of most copyright and licensing restrictions (Suber 2015). Some of this material will be held by digital repositories, but the open access movement is more about actively encouraging researchers to publish their material in an openly accessible format than trying to make older collections accessible in the first place. The question of open access however falls outside the scope of this research.

The other term used within the research question is the term 'digital environment'. This term is understood by some to relate to an environment where digital devices communicate and manage the content and activities within it (Kulesz 2017). The most popular example of a digital environment is the internet, which is a set of connected computers communicating with each other to form the infrastructure on which the World Wide Web sits. However, this is just one example of a digital environment. It can also refer to the digital environment found within an organisation. This is the infrastructure of devices that an organisation itself has implemented, for example for storage of material, or their own connected working environment (Kulesz 2017). It is also possible to talk of the digital environment in more expansive terms. For example, the European Union uses this term in their campaigns to create digital citizens across the union (The European Commission 2019) and a recent book publication of Boczkowski and Mitchelstein (2021) refers to the digital environment in similar terms (Boczkowski and Mitchelstein 2021). For this work the digital environment that is being referred to is the internet, as this is where most repositories are imagining to be making material accessible on. This environment is referred to on occasion as 'online public space.' The internet was chosen over the World Wide Web as a term, as some organisations may be using private connections to make material available to users.

Another term frequently used in this thesis is that of digital material. This refers to all material that is stored and given access to in a digital manner. This is not limited to only born-digital material, such as Word documents, 3D models, computer-aided design (CAD) drawings (which are detailed 2D or 3D illustrations), or digital photographs, but also includes digitised material or digital surrogates. This is because these digital surrogates can be viewed as their own enriched and useful entity, even when having an analogue equivalent (Nicholson 2013).

This research was not conducted in isolation and has been influenced by several factors. Firstly, the main research question and the reality of the disconnect between what is happening in practice and what is talked about is not new to me. Whilst working at the Archaeology Data Service (ADS), many of the constraints around providing access to digital material became apparent to me. The many frustrations around understanding users and providing access to this material in a meaningful way were part of my day-to-day practice, especially for newer formats of material that needed to be preserved, such as 3D-models or GIS maps. The work conducted at the ADS has therefore influenced the main research question around how this material should be made accessible.

This work has also been influenced by its specifically collaborative nature. This PhD is funded by the Arts and Humanities Research Council's London Arts and Humanities Partnership programme and in line with their guidance it is encouraged to collaborate with your partner institution, which in my case is TNA. I have done this extensively throughout my PhD, working with different teams at TNA including the digital preservation team, the research team, and the web archiving team. All details on these collaborative projects and how they have influenced my methods and subsequently my research can be found in the Methodology chapter (Chapter 2).

1.2 Contributions

The thesis explores how digital preservation practitioners are provisioning access in the digital environment. To provide context for this question, it sets out an overview of both how access to digital material has been provided in the past, and of the current discussions and thinking on this subject (Chapter 3). Such an overview has not been drawn up before and its creation is therefore an important contribution in its own right. Another contribution is the exploration of the digital preservation community. This thesis not only includes an outline of the community itself, but also of how it emerged, leading to a better understanding of how, and why, digital preservation practitioners currently think about their tasks and responsibilities (Chapter 4). The relevant literature is then expanded upon through a number of interviews with individuals within the digital preservation practice space as well as within adjacent spaces (Chapter 5). The aim of this work is to provide a clear overview on the provision of access and the different perspectives that this may take, therefore making it possible for digital preservation practitioners to more easily engage with access to digital material.

Additionally, the research is heavily influenced by the collaborative nature of this work. This has provided the opportunity to speak with practitioners on a regular basis, but also try out a number of concepts and ideas in practice. Examples of this are the Machine Learning Club that was run at TNA (Bell and Talboom 2022), the Humanities and Data Science Group with the Alan Turing institute (The Alan Turing Institute 2021) and extensive work completed on the provision of access to the UK Government Web Archive (UKGWA) (Storrar and Talboom 2019; Beavan et al. 2021). More detail on these projects can be found in Chapter 2.

1.3 Chapter Overview

This thesis follows a slightly different structure than would normally be expected. Instead of starting with a literature review, outlining the methodology, presenting the findings and discussing them, this thesis instead follows the different iterative research phases that were gone through before presenting the conclusion.

To further elaborate this structure **Chapter 2** outlines the research methods used which goes into detail on the collaborative nature of this work and highlights the iterative process used to conduct the research, it also gives an overview of the different interviewees who participated in it. After the methods chapter, three discussion chapters follow. All of these chapters have their own relevant literature, outline of context, presentation of findings, discussion and remarks looking forward. This presentation more accurately reflects and represents the iterative process by which conclusions were drawn and is thereby easier to follow and more transparent to the reader.

Chapter 3 focuses on the provision of access, highlighting the different technologies that have enabled or constrained access and setting this provision within its historical context. This chapter highlights a difference seen within the participants, with some interviewees able to clearly define the scope of their activities and others feeling more unsure and stuck. This difference is further explored in **Chapter 4**, which focuses on the digital preservation community. This chapter establishes a context for the digital preservation community by examining how it has emerged and who it is who identifies as a digital preservation practitioner. It then goes on to explore how the participants talk about the responsibilities they feel this identification brings, surfacing that, whilst these responsibilities are felt differently, many feel that that it does bring a wider societal responsibility or mission.

Chapter 5 focuses on the digital environment and on external factors that have an impact when working in this environment. Within this chapter a comparison is set up, by introducing voices from beyond digital preservation practice in the form of data journalists. Data

journalists also operate in the online public space with a societal mission and their perspective provides a fruitful counterpoint to further enhance understanding of that of digital preservation practice. **Chapter 6** concludes this work and highlights all findings from the previous chapters and answers the research question proposed above and highlights a number of future research directions.

2 Methodology

This research has a multi-phase iterative research design which reflects on the direction of the research in each phase. The research was organised in three stages, which are discussed below in detail. All three phases involved several interviews with relevant participants. Alongside interviews in each of the three phases relevant literature was reviewed. This approach made it possible to set the participants' perspectives within broader contexts and enabled a deeper understanding of where they were coming from and the ways their thinking might have been influenced and shaped.

Literature has therefore not been reviewed in what might be a typical way or form but rather it has been used as a methodological tool as outlined in more detail below. This chapter begins by discussing the general methodological approach, then gives details about the research phases and concludes with details on methods used during the research. Constraints and ethical considerations are discussed at the end of this chapter.

2.1 General methodological approach

One of the first approaches that I considered and ultimately rejected was a practice-based one. With a practice-based approach the emphasis lies on creating something, which results in a product at the end. This product does not necessarily have to be something tangible, but the focus of the research is on the process of creation (Candy 2006). As it was desired to focus this research differently; on understanding the negotiation of the provision of access by digital preservation practitioners, this approach did not seem suitable.

Another approach considered was a practice-led one, also called action research, where the researcher identifies a constraint, and then goes out in practice and collaboratively works to try and resolve the identified constraint in the field. This approach was not considered suitable for this work. Having come from a practice background, it seemed more suitable to actually take a step back and reflect on the current processes and constraints rather than continuing to work in practice. Therefore this work cannot be deemed as action research, as action research implies that there is a collaborative problem-solving relationship between the researcher and the interviewees (Pickard 2013, 157), which is not the case. Rather this work follows a more traditional research approach, where findings and theories are used to set up a number of recommendations for future action.

Thematic analysis and Grounded Theory seemed most suitable for this research. These two approaches are quite similar in their process, with coding and the discovery of broader patterns in the data being characteristics of both methods. However, Grounded Theory is

focused on finding one emerging theme from the material, whereas thematic analysis is more about emerging categories, or themes as the name would suggest. Another big difference between the two approaches is that Grounded Theory does not only define the research method, but also the underlying methodology (Charmaz 2006), whereas thematic analysis is not tied to a specific methodology and is, therefore, more theoretically flexible (Clarke and Braun 2017, 297; Braun and Clarke 2020) in that it can be combined with many different methodologies. This does mean that similar results can be reached from thematic analysis and Grounded Theory when the methodology is similar. For example, if a constructivist view is chosen when applying thematic analysis, a constructivist Grounded Theory approach may lead to similar conclusions (Charmaz 2008; Braun and Clarke 2006, 80–81).

This work adopted a thematic analysis approach. The reason why this approach was chosen over Grounded Theory was that thematic analysis offers more room for flexibility and is not tied to certain processes, such as the code book, leaving room for creativity when analysing and working with the gathered material. Additionally, this method is recommended for researchers early on in their career as it teaches a number of core skills which can then later be applied when conducting other forms of qualitative analysis (Braun and Clarke 2006, 78). Furthermore, something that is really crucial to this research, thematic analysis is seen as a useful approach when exploring the ways in which participants make meaning out of their experiences and how these experiences are informed by the world around them (Evans 2007). As the research involves both seeking out the experiences and perspectives of digital preservation practitioners as well as personal reflection through the lens of these perspectives and the reviewed literature, this seemed to be a fitting approach.

When looking at thematic analysis in more detail, Braun and Clarke (2019), who introduced the approach, make a distinction between coding reliability thematic analysis (TA), codebook TA and reflexive TA. Coding reliability TA comes from a positivist perspective where researchers are taught to code in a similar manner with a pre-determined codebook or coding frame. Codebook TA is similar to this approach, where a codebook is used, but is not necessarily grounded in a positivist view. Reflexive TA is a far more organic approach; not assuming that themes just emerge from the data, but rather that they are generated by the researcher by reflecting and comparing their material. The researcher actively creates the codes, they do not passively emerge from the data (Braun and Clarke 2019, 593–94).

Reflexive TA, an approach introduced by Braun and Clarke (2019), was followed for this research. The emphasis in this approach does not fall on the coding itself or on following procedures in the right manner, rather it prioritises on one's work and thoughtful engagement with the generated material (Braun and Clarke 2019). Reflexive TA suits this work as it has been heavily influenced by previous work and studies and, therefore, phases of reflection are important to understand where themes are emerging from and why. The reflexive TA approach goes well with the postpositivist view that this work has taken, in particular critical realism. This view has an understanding that here is a one social reality, but that there is an 'observable' world that is constructed by one's own perspectives and experiences (Pickard 2013, 7; Wiltshire and Ronkainen 2021). This fits very much into the idea that research is not objective, which is fitting in this case with it being heavily influenced by my background in the digital preservation field, and the nature of the collaboration with TNA. During the research, it was important for the researcher to reflect with the help of memoing in order to understand their subjectivity and where their thoughts and ideas were coming from.

The research method used in this project was interviews. Other research methods were considered, including working groups and surveys, but as an important part of this research was to explore how digital preservation practitioners perceive and feel about their role and responsibilities it was felt that other methods would not give them the room to explain and talk about their challenges and choices in detail. Case studies were considered as a potential research method, but as case studies give the possibility to go in-depth on a smaller number of participants, it was seen as too limiting for this study as there are many ways that access to different types of digital material can be provided.

As interviews lead to qualitative material, a number of qualitative research approaches were considered. This work is highly influenced by the collaboration with TNA but also by my experience and background in digital preservation. I have first person experience with making digital material available to users and understand many of the struggles faced from the reviewed literature and by co-workers around me. My background and the collaborative opportunities provided during my PhD influenced the work that I do and therefore it is important to make sure a method is chosen which makes it possible to reflect on the work throughout. The related projects to this work are found in Section 2.3.5 of this chapter.

2.2 Research Phases

This research followed an iterative approach split into three phases. Details are outlined for every separate phase and study group below and details on the overall methods used are found in Section 2.3. The overall trajectory of these phases can be summarised as follows. Firstly, consulting with the status quo was seen as important, which was the digital preservation community. Secondly, it was important to look at the group that was disrupting that community and trying something different, by experimenting with providing access to material for computational methods. And thirdly, I decided to look at a group who seemed to have similar issues to the digital preservation practitioners and work in a comparable setting, which are the data journalists.

2.2.1 Digital Preservation Practitioners – Phase 1

The first phase of the research sought out the views of digital preservation practitioners with respect to their feelings about and experiences of providing access to digital material. Before starting this round of interviews, the literature was consulted and a number of emerging themes around access and thinking about access were established. These themes can be found discussed throughout the following chapters and were used as headings for the Topic Guide in this phase (Appendix 1). As this research phase focused on digital preservation practitioners, it was important for the interviewees and the institutions they worked for to be involved and engaged in preserving and providing access to digital material.

The selection process for this phase was based on existing contacts and contacts suggested by TNA. It was made clear that there was an importance in trying to pick as diverse a pool of institutions as possible and the focus was on London, due it being deemed a test phase in first instance. This meant that TNA was able to provide contacts for Transport for London, Wellcome and the British Library. I used my own contacts to interview the ADS. The only struggle was to find a community archive that was available for an interview and had made material accessible in some way. After contacting a number of community archives across London, the London Community Video Archive was the one to reply.

The interviewees selected for this phase came from a number of different institutions, some being nationally significant repositories with years of experience, while others being small community archives. This diversity in interviewees was a crucial factor to the research as it would help understand the field of digital preservation from different perspectives and give a general idea of what trends and feelings are similar across the sector. For the first phase of research, organisations of five different types were chosen: one corporate archive, one

privately run archive, one library, one community archive and one specialist archive. Some of these institutions have a broad audience, whilst others are more focused on one type of users. All have a history of working with born-digital material, some only having born-digital material, whilst others also having a history of making analogue material, or their digital surrogates, accessible. And all have different mandates for collecting this material. Table 1 outlines some general characteristics of the different interviewees and lists who has been interviewed within each organisation.

Table 1 *- An overview of participants from the first phase of research. The chosen columns highlight the diverse nature of the interviewees.*

Organisation	Material	Reason for preserving digital material	Funding	Type	Size of institution
Archaeology Data Service (ADS) – Ray Moore	Born-digital	To preserve digital material for re-use purposes in the archaeological sector	Depositors pay a fee to archive the material	Specialist archive	Medium, 10-15 employees, all dedicated to archiving the material
Wellcome Trust – Alexandra Eveleigh	Born-digital, digitised and analogue	To continue the legacy left by Wellcome in collecting material around medicine and pharmaceutical practices	Shares in the company	Sector archive, museum and library	Large, dedicated digital preservation team
British Library (BL) – Maureen Pennock	Born-digital, digitised and analogue	Preserving as it is a legal requirement	Grant-in-Aid from Government	National Library	Large, dedicated digital preservation team
London Community Video Archive (LCVA) – Tony Dowmunt	Born-digital and digitised	Preserving out of necessity. The formats of the material have become obsolete and people who worked on the material are passing away, but there is a contemporary importance to the material	None, won a Lottery Fund bid, but this has run out	Community archive	Small, 2-3 people
Transport for London (TfL) – Tamara Thornhill	Born-digital, digitised and analogue	Preserving what the business is doing, a justification on why processes were done in a certain way	Company keeps the archive running	Corporate archive	Small, 3 people for a company of over 1500 employees.

A topic guide was created to help the participants familiarise themselves with the potential topics that would be discussed during the interviews. Additionally, a list of potential interview questions was created to help the interviews along and to ensure no topics were missed. These were split out into three sections: acquiring material, preserving material and making material accessible. Interview questions were informed by the consulted literature, which gave an idea of what topics and questions were of importance. The main aim of the interviews was to talk about access to digital material, but the other two topics were considered to be closely related in that they could influence the way that access to the material was being provided. Furthermore, these interviews were used as a way to see if the claims made in the relevant literature were found to be true when talking to the digital preservation practitioners. Before diving into the topics, the participants were also asked to outline their role and background. The topic guide and initial set of interview questions can be found in Appendix 1. As the interviews were of a semi-structured nature, the interview questions were only used as a guide to ensure no specific topics were missed.

During the interviews a certain vocabulary was used. This was decided upon in part not to confuse participants who might not have been as familiar with terms, from the archiving community, that are used within digital preservation settings. As the research was not limited to the archiving sector but looked at participants from all across the digital preservation field, making it crucial to pick terms that would appeal to and be understandable by individuals from a wide range of backgrounds. Details on how the digital preservation community has emerged and who identifies as a digital preservation practitioner can be found in Chapter 4.

Five interviews were completed in this phase. Due to the process of constantly reflecting on the interviews and the transcribing process, it was apparent that saturation was reached quite quickly. Despite picking a broad range of institutions with different backgrounds, similar themes emerged in all interviews, which were similar to the ones picked up on when reviewing the literature. The first group of interviewees expressed a large amount of doubt mainly around providing access to digital material. One of the other themes that kept coming up was the uncertainty around automation and the emphasis on preservation.

This meant that for the next phase of research other individuals had to be consulted who thought differently about access. For this reason, a decision was made to focus, in the next phase, on interviewing individuals who were approaching the provision of access to digital material in terms of providing computational access (information on this approach can be

found in Chapter 3). This group was chosen to see if the same concerns were raised by people providing computational access and if this was a feasible way of providing access to the digital material.

Throughout the rest of the thesis, the group of people that were interviewed during the first phases of the research is referred to as Study Group 1.

2.2.2 Computational Access Providers – Phase 2

For the second phase of the research a different direction was decided upon. The focus of this phase was on the potential of computational access as a means of providing access to digital material. In short, computational access is providing access to digital material to make it possible to compute over this material; a detailed overview of this approach can be found in Chapter 3. The choice to focus on computational access was highly influenced by the work with TNA and in particular the project with the Alan Turing Institute. This showcased a different way of accessing this material that went beyond the digital catalogue that the UK Government Web Archive, based at TNA, makes their material available through (Beavan et al. 2021). My own experience as well with Natural Language Processing during my Master's highlighted that their were different ways to explore digital material as data.

To identify individuals working with digital material in this way, relevant literature was consulted. This involved identifying projects and individuals within the digital preservation community who were working on seeing digital archival material as data. Identifying interviewees was made easier by help from co-workers at TNA and other acquaintances in the field, as they were able to identify individuals taking this novel approach to access.

The main criterion for selecting the participants in this phase was that they were making digital material accessible in a computational manner. This could either mean that material was made available as datasets, but platforms and APIs were also considered. Furthermore, making the material available as data did not have to be fully implemented at the time of the interview, but at the bare minimum there needed to be a project in place experimenting or thinking about this type of access. As with Study Group 1 a topic guide was created and circulated to the participants before interviews took place. This topic guide contained slightly different topics than that for the first group and focused on the type of access being provided and how this was managed within the organisation. This meant that three topics were of interest: the material being made accessible, the framework that this material was made accessible through, and the users of this framework/material. The interview questions reflected these three topics in structure, but again the interviews were started by asking the

interviewees about their background and role within the organisation. The topic guide and interview questions for this phase can be found in Appendix 1.

Because of the time when the interviews were conducted, which was during the Covid-19 pandemic, all interviews were conducted virtually. This led to limitations as the organisations where participants worked were not visited; details on this impact can be found in Section 2.4. Additionally, the pandemic opened up the possibility of interviewing participants outside of the initial proposed geographical location of the United Kingdom and the Netherlands. Therefore, participants from outside of this geographical location could be contacted. Pre-arranged interviews were already set up, but it opened the possibility to interview GLAM Workbench, Archives Unleashed and HathiTrust, all projects and organisations that have done extensive work in making their material available as data. Again, it was ensured that the vocabulary would be understood by the interviewees.

Making the choice to go outside of the original scope opened up the possibility to see if similar problems were felt in institutions with different copyright and legislation in place than only the UK and the Netherlands. As can be seen from the guidance produced by the DPC, digital preservation is a global issue. It should be kept in mind here that it could have caused issues in comparing the material collected, but as similar processes and approaches were in place, this was deemed at a minimum.

The participants in this phase of the research are from a diverse mix of backgrounds and can be roughly split into three different groups, all providing or envisioning computational access for their digitally preserved material. The first group are technical people, active within the digital preservation field; these could be product managers or technical leads on projects. These participants do not have a humanities background but do have an understanding of the work carried out by humanities scholars. The second group is very similar to Study Group 1, i.e. participants who identify with the digital preservation field. The big difference between the people in this group and Study Group 1 is that the former are trying to provide a more computational way of accessing the material. Then there is a third group that floats between these two groups; these are participants with a humanities background that have enhanced their computational skills to enable themselves to engage more in the technical side of this work.

Not only the background of these participants is more diverse than the previous group, but also the way in which the material is preserved is of interest here. Not every single participant in this group preserves the material themselves. As the criteria for this phase was not

necessarily preserving the material itself, as the focus was on access, the possibility of interviewing a broader range of people opened up, including those who were involved with projects such as CLARIAH (Melgar-Estrada et al. 2019), GLAM Workbench (Sherratt 2020) and Archives Unleashed (Ruest et al. 2020). None of these projects hold material themselves, but instead create tools or platforms to enable access to digitally preserved material from other organisations.

To ensure that as wide a group of participants as possible was interviewed, the size of the organisations was taken into consideration again. As with Study Group 1, some of these organisations have long been established and acknowledged for their digital preservation work, whilst others are smaller organisations that may not have been doing this for as long. Furthermore, some of the interviews in this group had several participants attending one interview. From this point onwards this phase of research is identified as Study Group 2.

Table 2 - *An overview of participants from the second phase of research. This last column in this table highlights the computational access approach taken by the interviewed party and the reason they were interviewed in the first place.*

Organisation/Project	Background	Interviewee(s)/Role(s)	Computational Access Approach
CLARIAH Project	**Technical**	**Roeland Ordelman** – Product Manager	Has a platform for access with an underlying API
Legislation.gov.uk	**Technical**	**Tamara Izzo** – Legislation Data Analyst	User interface for individual downloads, with an API for bulk access
GLAM Workbench	**Humanities**	**Tim Sherrat** – Project leader	Makes a set of Jupyter Notebooks available to explain how computational access can be done
Archives Unleashed	**Humanities**	**Ian Milligan** – Principal Investigator	Set of tools to help with the access of web files, both for users and archivists alike

Wellcome Trust	**Technical**	**Jonathan Tweed** – Technical Product Manager	Catalogue with an underlying API
British Library	**Humanities**	**Rachel Foss** – Curator of Modern Literary Manuscripts; **Jonathan Pledge** – Digital Archivist; **Callum McKean** – Digital Archivist	Still thinking about this type of access, number of experimentations have been done
Koninklijke Bibliotheek	**Humanities**	**Steven Claeyssens** – Digital curator	Still thinking about this type of access, but number of experimentations have been done
Project Alpha (TNA)	**Technical**	**Technical Lead**	New way of thinking at TNA, API first
UK Data Service	**Social sciences**	**Louise Corti** – Head of Secure Research Service Development	Datasets are available with an environment for more secure data
HathiTrust	**Humanities**	**Digital Scholarship Librarian**; **Graham Dethmers -** Metadata Analyst	Datasets, catalogue and secure digital environment
International Institute of Social History (IISG)	**Humanities**	**Eric De Ruijter** - Manager Collections; **Robert Gillesse** – Digital Archivist	Catalogue, datasets with the help of linked data

After completing eleven interviews during this phase saturation was reached. Saturation was acknowledged after memoing and reflecting upon the research after completing each interview. This was acknowledged when participants started repeating themes around the importance of a digital infrastructure and their frustration with communicating between technical individuals and digital preservation practitioners. This was also the stage when it became clear that two distinct groups were being interviewed. These groups were first seen as being divided into those people providing access to their collections as data and those who were not, but the distinction then became more nuanced - between people who

strongly identify as digital preservation practitioners and those of a more technical mindset working in this space.

Study Group 2 was able to provide insight into the future directions of digital preservation and access to digital material. When working on this phase it was acknowledged that several other professions are also experiencing a digital shift and therefore the decision was made that it would be beneficial to gain insight from another profession; this would be the final phase, Phase 3, which is discussed below.

2.2.3 Data Journalists – Phase 3

The third phase of research was conducted with participants from outside the field of digital preservation, who have nonetheless gone through a similar digital shift. For this phase data journalists were chosen. This decision was made after being introduced to data journalism, and more specifically, investigative journalism, when attending the keynote at iPres 2019 (Higgins 2019). Data journalists provide access to the material that supports their articles, and at the same time they have similar concerns (to digital preservation practitioners) with being trusted by the public. This group was chosen over other similar fields such as the research data managers or corporate data managers as the data journalists have a different relationship with this material and are already making it available in articles within in the online public space. As the online public space is of specific interest when looking at access to digital material within this work, this field was seen as a group where useful insights may be gained.

To be able to reflect and understand where this profession was coming from, relevant literature was consulted. The overview of this review can be found in Chapter 5, as this is the chapter that introduces the data journalists into the research. Essentially it comes down to data journalists being seen as a fascinating discipline, as their core values around trust and context are similar to the digital preservation practitioners'. They may not be preserving digital material, but they are publishing material in the same online public space as digital preservation practitioners. Therefore, it is of interest to compare their approach and see what digital preservation practitioners could learn from them and vice versa.

Approaching participants was significantly more difficult than in previous rounds. There was a struggle in pinpointing exact interviewees to talk to, but also in establishing the initial contact. It was ensured that the topic guide for this round was as clear as possible. The interview questions for this round were very similar in structure to the interview questions for Study Group 1, but instead focused on material collected by data journalists. Again, it was

ensured that jargon was not used and in cases where jargon could not be avoided a detailed description was provided. The interview questions and topic guide can be found in Appendix 1. When researching the field of data journalism, it was acknowledged that there were three different categories of interest. The first being data journalists active in traditional news organisations, the second being data journalists active in newly established organisations and the third being data journalists teaching the profession. All three of these categories were initially covered, as can be seen in Table 3.

Table 3 - *An overview of participants from the third phase of research. The last column highlights the specific area of data journalism that the participants are active in.*

Organisation/Project	Interviewee/Role	Data Journalism
Cardiff University	**Aidan O'Donnell –** Lecturer Data Journalism	Data Journalism program, longest established one in the UK
Bellingcat	**Eliot Higgins –** Investigative Journalist	Investigative journalism platform, established in 2014
BBC	**Jeremy Tarling –** Lead Data Governance Specialist	News organisation
Full Fact	**Leo Benedictus –** Fact Checker	Fact checking organisation, established in 2010

In this research phase the goal was not to reach saturation, but instead to gain insights and understanding to be able to compare the practice of data journalism with the digital preservation practice. The third round of interviews with the data journalists, will be referred to as Study Group 3 from this point on in the work. The insights gained from talking to them are discussed in Chapter 5. This phase concludes the data collection process and will lead to the creation of themes and the initial writing of the analysis of the research.

2.3 Methods used during the research

Throughout the three phases outlined above the core approach and methods used remained the same, to ensure that the data collected from the different phases would be consistent enough to allow comparison and common analysis. Below these methods are discussed. This includes the method employed to identify the relevant literature, methods used for conducting interviews, and an outline of how material was analysed.

2.3.1 Interviewing

Across all three phases, the interviews conducted followed a similar pattern which is discussed in this part. All the interviews took a semi-structured form. This approach was chosen as it gives time to prepare for the interviews with a number of topics to be discussed, but the interviews are flexible enough to allow the respondent to answer the questions in their own terms and discuss topics of their own interest (Choak 2012). It is possible to use the themes and topics identified from the literature as a starting point for conversation, but enough room is left for the interviewees to elaborate on any processes or parts that they deem important.

For all interviews a topic guide was created; specific topics relating to the different research phases are found in Section 2.2. These topics were used to create a number of interview questions. It was not deemed necessary to ask all research questions in the exact nature formulated in these interview questions, instead these questions were used as a guide to ensure that all relevant topics were discussed during the interviews. The topic guides were circulated with the interviewees before the interview to make it possible for them to familiarise themselves with the topics to be discussed.

Participants were chosen based on criteria outlined during each phase. Colleagues at TNA were a big help when making the initial contact with participants as a number of them were known by the staff or used to work at TNA. The snowball sampling was used during the interviewing stage to find other potential interviewees. This is a technique where participants are asked to help identify other potential interviewees or projects of interest (Etikan 2016).

It was always ensured that enough time was taken between each interview. Interviews were not completed in bulk but spaced out with preferably two weeks between the interviews. This was done to make it possible to transcribe and reflect upon each interview before starting the next one. For the first two phases, interviewing was completed the moment that saturation was reached. For the third phase interviews were completed when enough comparative material had been collected.

Before the interviews started the participants were asked to sign a consent form, setting out if they understood the requirements. When this initial contact was made the topic guide was provided to make it possible for participants to orientate themselves before the interview was conducted. Interviewees were asked if they wanted to be anonymised, pseudo-anonymised or identifiable. The option to be identifiable was included for the participants as

the field of digital preservation is quite small and even with being anonymised, there was still a chance of colleagues, or supervisors, being able to identify the participants. If the participant agreed, the interviews were recorded on two devices and during the interviews notes were taken. The interviews were between 45-90 minutes long.

The choice of using two recording devices was made as it would always ensure a backup of the data was available. This turned out to be a useful approach when transcribing, as some of the audio was inaudible on the first device and sometimes the second audio recording was clearer and therefore easier to transcribe.

The first phase of interviews was conducted in the summer of 2019 and most of these interviews were conducted in person at the participants' institutions. This was no longer possible for the second and third phase of the research, conducted in the second half of 2020. Due to the pandemic, it was impossible to visit participants at their organisations. Further details on the impact of the pandemic can be found in the limitations section of this chapter. This meant that all interviews during the data collection were conducted online during Phase 2 and 3. To ensure an encrypted service was used, as preferred by the Ethics committee at UCL, Jitsi and Microsoft Teams were used. Later Zoom was additionally approved by UCL for conducting interviews, as there was a number of security concerns around this software at the time. The recording of these interviews was not completed with the built-in recording option that the online platforms offer, as it was unclear if built-in recording tools were in line with GDPR.

This impacted data collection slightly. Due to the pandemic-related restrictions, the ethics form had to be updated to outline that all pending interviews would be conducted online. This caused some delay, but also opened up the possibility to interview participants outside of the initially approved regions, which were the United Kingdom and the Netherlands. It was now possible to approach anyone anywhere in the world and meant that the pool of participants expanded. Further details on this can be found in the limitations section of this chapter.

2.3.2 Transcribing and Coding

Transcribing of the interviews was performed directly after completing the interviews. Transcribing was done manually and was not outsourced or completed by any software. This approach was chosen to help with familiarising with the data, which is seen as beneficial when using a TA approach (Braun and Clarke 2006, 87–88). After the transcription was complete the interviews were returned to the interviewees for checking. This check enabled

the participants to omit or clarify anything said during the interview. After checking was complete, the original interview recording was deleted. All transcriptions were written in Word. An example section of a transcription can be found in the Appendix 2.

After all interviews and transcriptions were completed in one phase, an initial round of coding was conducted. During this initial round, preliminary jottings were made in the margins of the transcriptions (Figure 1). This was done manually to get used to the coding process. After these jottings were made, the first round of coding was completed. The approach of holistic coding was used, as this is a recommended method for first-time coders (Saldaña 2009, 118–19). These codes were initially done manually, but later were imported into NVivo (Figure 2), which is a qualitative data analysis computer software package. NVivo is recommended, as the tools within the software offer flexibility when analysing a large data corpus, making it easy to change and enhance codes throughout the process (Richards 1999).

This approach to coding was conducted for every phase of the research and another round of coding was completed after all the data was collected. The final themes from this work are around the provision of access, the role of the digital preservation practitioners and the society and the digital. These themes will further be reflected on in the following chapters, details on how saturation was reached and the emergence of these themes can be found in Section 2.2.

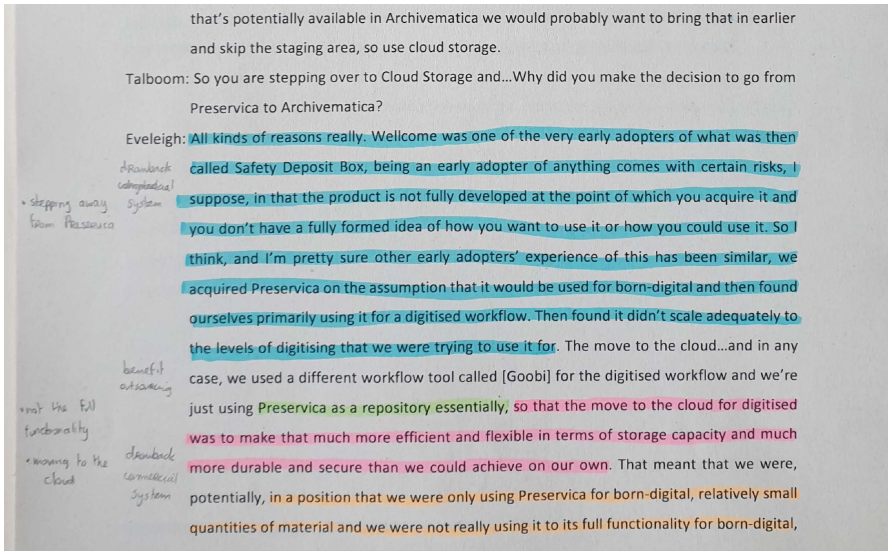

Figure 1 *- Preliminary jottings in the margins of a transcription and the first codes done by hand.*

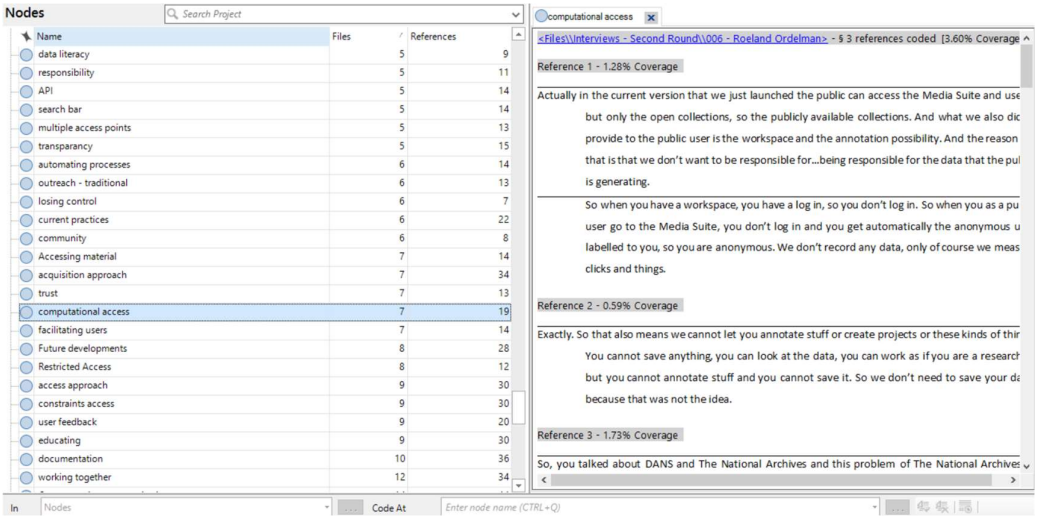

Figure 2 - *Example created nodes in NVivo. The left shows an overview of the nodes, the right is a highlight of the 'computational access' node, highlighted where these codes can be found in the interview data.*

2.3.3 Memoing

Memoing was done throughout this work, when conducting interviews, but also when consulting the literature or when any other ideas or concepts emerged. Memoing is predominantly a technique used within Grounded Theory (Saldaña 2009, 32–44), but as discussed above, there is an overlap between the methods used in Grounded Theory and TA. Memoing may not be named under the same term in Braun and Clarke's (2006) article, but they do place an emphasis on the importance of writing and reflecting on the work at every stage of the research (Braun and Clarke 2006, 86). As reflexive TA encourages researchers to be as reflexive as possible, memoing seemed to provide an easy and structured way to approach reflective writing. Additionally, the use of memos was very helpful in sorting through and structuring the collected literature.

Memos were written throughout the full process of coding and this was conducted in a systematic order, before the interviews, after the interviews, after the transcriptions, after the preliminary jottings and after the first round of coding. However, more memos were made of thoughts and reflections or relevant literature whenever appropriate. The memos were all written electronically and named in a systematic way, starting with the date and a brief description of what the memo contained. As the memos were created electronically, a full-text search could be used to locate the correct memo when needed. An example of a memo can be found in Appendix 2.

2.3.4 Reviewing the literature

The collected literature for this research expands further than academically published articles and books. As a large part of the work and practice of the digital preservation field does not necessarily happen in an academic setting it was important to ensure that work outside of this academic scope was consulted alongside the academic works. Other material that was included consisted of blog posts, reports and conference proceedings, but also institutional reports or announcements.

To locate relevant material the first step consisted of analysing a number of digital strategies published by institutions looking after digital material in the UK and the Netherlands. Examples of this are the digital strategies published by TNA, British Library and the Network for Digital Cultural Heritage in the Netherlands (The National Archives 2017b; The British Library 2017; Digital Heritage Network 2019). The references and discussed topics within these digital strategies were then used as a starting point for further investigation to locate relevant material. This was done by following up the references or doing more research on mentioned topics, such as Web 2.0 and the OAIS model. Furthermore, information or relevant topics mentioned by colleagues at TNA were followed up and added to the literature review where relevant. Lastly, the Library and Information Science Abstract (LISA) database, Google Scholar and the UCL catalogue were used to locate any further relevant material in these themes.

The keywords used during these searches were based on topics highlighted in the digital strategies and the searches were conducted to ensure that all relevant material had been identified. Examples of these keywords were digital preservation, digital archiving, designated community and born-digital material. The first round of the literature review was completed before starting the interviews but was reflected and updated throughout the research, as pointed out in the research phases sections (Section 2.2). It should be noted here that the literature review did not start from an academic reference source, as it was noted from the start of this work that a lot of the reviewed material, such as the digital strategies, will not show up in these more traditional academic libraries. However, the LISA database, Google Scholar and UCL catalogue were checked with relevant terms to ensure nothing important was missed.

This work does not have a traditional literature review chapter, instead the decision was made to summarise the relevant material at the start of all the following discussion chapters. This decision was made as the literature review was not completed in one sitting at the start of the work; and was done in phases when reflecting on collected material or before starting

a new research phase. Having the relevant literature in the discussion chapters therefore reflects better on the research approach that was followed and helps the reader follow along in a more meaningful way than summarising and reflecting on all literature in one single chapter.

2.3.5 Doing

This research was not conducted in isolation and has been influenced by several factors. First off, the main research question and frustration around providing access is not new to me. My background lies in computer science and archaeology where I worked as a data manager on an archaeological project, sparking my interest in storing and preserving digitally generated archaeological material. This led to me doing a Master's in Archaeological Information Systems where I was fortunate enough to complete a work placement with the ADS, a digital archive focusing on preserving and making accessible digital material from within the cultural heritage sector.

My master's project was on the use of Natural Language Processing to increase the discoverability of zooarchaeological terms in archaeological reports, which were part of the ADS's collection and currently only searchable through manually inputted keywords. After finishing this project, I continued working as a digital archivist at the ADS full time. Here I tried to continue working on increasing the discoverability of the archaeological reports that I started working on during my master's and did my best to implement components of this work into the actual infrastructure of the ADS, which did not work due to numerous technical and managerial issues. This is where I started to understand the possibilities that digital material, and specifically born-digital material, may have, but also the constraints in making it accessible. This experience has highly influenced my current research topic, as I can see the possibilities, just like many other digital preservation practitioners can probably see across the field, but also understand the struggles of making this digital material more accessible for users of the digital repository.

Another big influence of my thesis is that this is a collaborative PhD, meaning that in line with the Arts and Humanities Research Council's London Arts and Humanities Partnership programme students are encouraged to collaborate with their partner institutions, in my case TNA. There are no strict rules on the form of this collaboration. I ended up doing a number of projects with colleagues across TNA which related to the accessibility of digital material. As the problem of access to digital material is a wider problem within the community it seemed important to view this from a larger perspective than only TNA, especially considering that at the time this PhD commenced TNA has just started grappling

with the concepts and difficulties in providing access to digital material. Therefore, it was deemed more useful to talk to individuals that were trying different things within the field.

This has resulted in a number of recommendations in the conclusion of this thesis. However, the projects undertaken in collaboration with TNA throughout my PhD heavily influenced my work and helped to shape my research direction:

- **December 2018 – December 2020, Derbyshire Project –** Started a project with Ian Henderson on the Derbyshire collection that is held by TNA. Currently large parts of this collection are not accessible by the public, even with it being an open record. We looked at the original source material and generated a number of different options that could help users not only access but also navigate this large and complicated collection on the internet. To finish off the Derbyshire project we gave a talk at CityLIS summarising our ideas.
- **Summer 2019, Access for Whom? –** Simultaneously to doing a more hands-on project with the Derbyshire project. I worked with another colleague, David Underdown, on a more theoretical piece around the term Designated Community. This is a term that has emerged from the OAIS model and is part of many digital preservation certifications. The term references the group that institutions make their material accessible for. Further details on OAIS can be found in Chapter 4. At the time of researching the model we saw numerous digital preservation practitioners struggling with Designated Community term, as the audience that accesses material is different on the internet than for the analogue material which is only accessible in the reading room. Within this piece of work we explored the term and how it is seen as a box-ticking experiment for many institutions. For TNA this term is problematic as they state that they make their material available to everyone, which becomes impossible within a digital space like the internet, you cannot cater for everyone. In the findings of the work we propose to not split user groups by certain traits or professions (e.g. public user, researcher, etc) but to look at the digital skills of the users instead: non-digital, digitally curious and digital. Hopefully making it easier to cater towards different user groups within the archive. This way of thinking really shaped my earlier frustrations with the OAIS model and helped me work out what the model was able to bring to the discipline, but also where it was lacking. This work was summarised in a blog post for the Digital Preservation Coalition (DPC) (Talboom and Underdown 2019).

- **Summer 2019 – Summer 2020, Machine Learning Club –** At the beginning of my PhD I was still partly working on Natural Language Processing tool I worked on during my master's and ended up giving an internal talk at TNA in this work. This is where I met Mark Bell and we realised that both of us were seeing an increase in interest for computational methods, and specifically machine learning techniques within TNA. But we also realised the frustration that many colleagues had when learning about machine learning, as all the training material and courses are aimed towards computer scientists or technical people with no examples for the digital preservation community. We decided to start running a Machine Learning Club where we explained basic concepts of Machine Learning and used Google Colab, which is an browser-run environment that combines executable code and text in one place (Google 2020) This offered the attendees the possibility to play around with some examples themselves. All these examples were set up with archival material, so all of them were relevant to colleagues at TNA. The aim was for us to give our colleagues more confidence when talking about computational concepts with more technical people and not turn them into data scientists. Our experience in running this club have been summarised in a book chapter (Bell and Talboom 2022). Running the Machine Learning Club help me shape the ideas around what skills digital preservation practitioners need, which is more around being able to communicate technical ideas than to implement these ideas themselves.
- **Summer 2019 – Spring 2022 Numerous projects on the Web Archive** – The UK Government Web Archive (UKGWA) which is managed by TNA is one of the only examples of an openly accessible web archive in the online public space of the internet. This makes it a great resource to explore new and interesting concepts to make born-digital material accessible to users. I did a number of projects that used the UKGWA as their main resource. Starting with a workshop run by the Computational Archival Science (CAS) workgroup where we explored the possibility of graph networks to visualise the UKGWA. Tom Storrar and I outlined some of our insights in a blog post that can be found on the TNA website (Storrar and Talboom 2019). At the end of 2019 I applied for the Data Study Group at the Alan Turning Institute, where again the source material ended up coming from the UKGWA. This time exploring the possibility of using topic modelling to make it easier to discover material within the web archive, a white paper has been published outlining the findings of this study group (Beavan et al. 2021). In the summer of 2021 Mark Bell

and I decided to do more work around the Google Colab that we worked on for the Machine Learning Club, this time focusing on trying to showcase a different way into the UKGWA that went beyond the current interface. This is in line with the collections as data, or computational access concept, which will be highlighted in Chapter 3. The notebooks were inspired by the notebooks hosted by GLAM Workbench on archived web material (Sherratt 2021). But we also decided to highlight some of the unique concepts of the UKGWA, as it is catalogued within TNA's Discovery catalogue. This work with Mark Bell is currently being turned into an article. Working with the born-digital material of the UKGWA gave me some hands-on experience of what computational access may look like in practice, and what constraints and possibilities this approach is able to provide.

- **Spring 2020 – Summer 2021, Humanities & Data Science Discussion Group –** During the Data Study Group at the Alan Turing institute Federico Nanni and I talked about the frustration of communicating ideas across different disciplines. This led to us starting a discussion group that invited people from the humanities and data science to do just that, to discuss topics that are of interest and are current in both disciplines but are discussed separately. We kept notes on HackMD and discussed topics such as open-source journalism, non-English Natural Language Processing and ethical implications of archiving the web (The Alan Turing Institute 2021). This discussion group has helped me shape numerous ideas around how different disciplines think about similar topics that are important to the humanities and how they are communicated across disciplines.
- **Autumn 2021- Summer 2022, SSI Fellowship –** This project will still be running while I complete my PhD. As will become apparent in the Chapter 3, institutions are starting to experiment with making collections available as data. However, there is a lack of resources for digital preservation practitioners in making the first steps towards providing this type of access. I successfully applied for funding for a Software Sustainability Institute (SSI) fellowship to have funds to work with the DPC on a guide to aid the digital preservation practitioners to take their first steps towards computational access (Talboom and Digital Preservation Coalition 2022). Alongside this guide a number of workshops (Talboom 2022) and events (Digital Preservation Coalition 2022b) will be run, which will highlight the main concepts of the guide, but also provide examples and case studies. Additionally, the guide will be a starting point for discussion during a panel at iPres 2022.

2.4 Ethics & Limitations

An ethics application was submitted to the ethics committee at UCL under number 15243/001 and was approved on the 3rd of May 2019. This form was then updated at the start of the pandemic informing the committee that all interviews would move online, which was approved by the committee on the 27th of March 2020.

Within the ethics form a number of ethical issues were pointed out. An important issue was that, given the size of the digital preservation community, which is small, it was possible that participants would want to self-sensor to some extent with regard to what they would say during the interview, as they would be aware that colleagues and employers might read the results. This was why all transcriptions were checked by participants after the interviews, to ensure that they could censor or update anything that was said. It was ensured that this issue was pointed out in the consent form and that it was made clear to participants, whether or not they chose to be named in the final thesis, that there was a strong chance that they would still be identifiable, especially by their employers.

Interviews were deemed to be the most appropriate way to collect data for this research, as discussed in the first section of this chapter, however this does come with a number of limitations. The number of participants that could be interviewed was limited. I tried to ensure as many different aspects and institutions were covered, but due to the nature of interviews this cannot be guaranteed as much as for example with surveys where a larger pool of participants can be gathered.

When conducting the interviews, the participants have the choice to stay anonymous, pseudo-anonymous or be identifiable. As the field of digital preservation is relatively small, even when being anonymised there is still a possibility that the participants may be recognised, this can have an impact on the interviews, as participants may withhold information. As this is a collaborative research project with TNA, and TNA playing a leading role in the digital preservation and archiving sector, participants may have felt uncomfortable talking about certain information or aspects of their process during the interviews. Participants were always made aware of this connection before the interview was conducted.

For the first round of interviews, which were conducted pre-pandemic, a number of limitations were in place. One being that ethical approval was achieved for conducting in-person interviews in both the United Kingdom and the Netherlands, which both have an

active digital preservation community. This does mean a geographical limitation was reached, as it would potentially not cover all views of digital preservation.

For the second and third round of interviews, which were conducted during the pandemic a number of other limitations were in place. Because of all the interviews having to be conducted virtually a revision of the ethics form was made, which led to a slight delay in the work. Moving all interviews to an online space both had its advantages and disadvantages. The advantage was that the geographical limitation that was in place during the first round of interviews was not an issue anymore, as the interviews could be opened up to anyone regardless of their location, making it possible to cover a wider variety of participants. However, all the interviews were conducted virtually, which meant that no site visits were made anymore. These site visits, which did happen during the first round of interviews, provided the possibility of gaining more context and made it easier to establish further contact for potential interviews. Furthermore, as pointed out by Pickard (2013), doing interviews online leads to difficulties in picking up on visual cues (Pickard 2013, 203). The pandemic impacted the duration of the data collection process; not only did some potential participants leave their positions during the pandemic, also the replies and setting up interview dates took substantially longer than during the first round of interviews.

3 Provision of access

To begin addressing the main research question of how digital preservation practice is negotiating the provision of access in the digital environment, this chapter examines the ways in which the provision of access has been and is currently being negotiated by the digital preservation community. As discussed in section 1.1, the assumption is that access should be provided in the online public space of the internet, as outlined by a number of people in the field as the best approach to be taken (Appel et al. 2015; The National Archives 2019; Stewart and Breitweiser 2019). This chapter outlines the evolution of engagement with the internet and how this has evolved through a number of different technologies. After providing this context it becomes possible to discuss different perspectives on the topic gathered during the interviews.

3.1 Web 1.0

In this chapter, the provision of access is discussed in relation to the evolution of the internet from Web 1.0 to Web 2.0 and Web 3.0 (Shivalingaiah and Naik 2008). Each of these eras is, to a certain extent, an evolution of the previous one, but they can also be seen as existing simultaneously. As an example, a repository could provide their users access to material through a static online catalogue where information is only provided by the repository themselves, this would be considered Web 1.0 but at the same time engage with their users on social media, which would be considered Web 2.0.

Before highlighting the responses of digital preservation practitioners to the different iterations of the web it should be highlighted that the online presence of digital repositories has only been sparsely documented (Anderson 2008; Bunn 2019). Attention tends to focus on the future and on new technologies, such that the history of online access is poorly documented or not documented at all, making it difficult to pinpoint exact dates. Within studies of digital libraries, archival research and virtual museums there have been only a few studies evaluating online presence (Costa and Silva 2012; Rahimi et al. 2018; Everstijn 2019). These studies tended to focus on the educational benefits and the interface or site structure, not on the actual evolution of these resources. The lack of study on this evolution makes the writing of this section challenging as the focus of the available literature is not on summarising or providing an overview of the past. By producing such an overview so therefore, this chapter provides a valuable contribution to the field and the sparsity of the existing literature has been supplemented by the consultation of web archives in which archived digital repository websites can be found.

One of the few studies which sets out the temporal evolution of archival resources offered online was conducted by Anderson (2008) who presented the Model for Archive Web Development (MAWD). In this model he compared archival web resources from July 2004 and September 2007 to devise six evolutionary categories, ranging from an archive publishing general information online, to interactive finding aids. Anderson concluded by stating that not much has changed in the period under study, but also that there was no decay perceived, as all sites were still accessible (Anderson 2008). This article is outdated now as it did not envision digital aids to be used further than to find material before going into a physical archival space to consult them, but it does give a small window into the evolution of sources for that period.

The online presence of archive services started with the emergence of the internet in the 1990s. The first iteration of the internet, Web 1.0, focused mainly on static websites which were used to publish static information which was institutionally owned and reliable (Shivalingaiah and Naik 2008). These first digital repository websites were mainly used as a promotional window which gave the user an overview of what could be consulted when visiting the physical location. An example of this is the 'Medieval Illuminated Manuscript' web pages created by the National Library of the Netherlands. This website, dating from around 2001, gave the user the possibility to browse a number of pages from a manuscript, but no full texts were available (Koninklijke Bibliotheek 2001).

In the next stage of evolution, online catalogues started to appear. Putting an exact date on these online catalogues is difficult, as memory institutions themselves have a lack of documentation regarding their online presence. Nevertheless, by using the Internet Archive's Wayback Machine, an approximate idea of when these were first released to the public can be made; some examples being The National Archives UK releasing Procat somewhere around 2001 (Public Record Office 2001; The National Archives 2017a), The Book Catalog released by The National Archives US somewhere around 2000 (National Archives and Records Administration 2000), and the KB-Catalogue released by The Dutch National Library somewhere around 2000 (Koninklijke Bibliotheek 2000). These online catalogues gave the user the opportunity to search through the metadata to see what the collections held in more detail. Instead of a small, curated set of material, it was now possible to gain information about most of what the memory institutions held. If the user found anything to their liking, it was possible to order this material on a case-by-case basis.

Mass digitisation started picking up in the early 2000s and an increasing number of digitised copies of original material became available through digital repository websites. This mass digitisation can be partly attributed to the growing commercial interest in such material by for-profit companies offering memory institutions the possibility to digitise collections that would otherwise not be financially feasible (Jensen 2020, 4–5), which Thylstrup argues is political (Thylstrup 2019). It became possible to not only order the analogue version of material through the catalogue, but in some cases also to consult a digitised copy of the material on the internet. These digital surrogates offered a more convenient way of accessing this material (Romein et al. 2020).

From the early 2010s a further shift can be seen. As more digitised copies of material were made available, new possibilities for digital material started to arise and the digitised copies went from being seen as digital surrogates, to being seen as enriched and useful data in their own right (Nicholson 2013). This shift was not only influenced by the growing amount of digitised material becoming available, but also the arrival of born-digital material. Born-digital material is slightly different from digitised material, as there is no analogue copy. Making born-digital material only viewable in some type of digital environment, either on a terminal in the reading room or the online public space. Pinpointing the exact data of this shift is difficult, although it can be concluded that the shift took place towards the close of the 2000s (Anderson 2008). There are archived websites, as mentioned, but it is impossible to consult these catalogues as interactive tools, as the current state of web capture technologies does not allow the full functionality of database-centric and dynamic websites to be maintained (Lohndorf 2022).

3.2 Web 2.0

The response of digital preservation practitioners to Web 1.0 was mainly a technical one. It was about getting a better understanding of the tools and establishing a presence online. However, this response changes with the next iteration of the web. In the early 2000s a new type of web emerges called Web 2.0, also known as the Participatory or Social Web. The term Web 2.0 was first mentioned by Darcy DiNucci in 1999 (DiNucci 1999) and was later popularised by Tim O'Reilly in 2005 (O'Reilly 2005). The idea of this era of the web is that users do not merely read websites but are also able to contribute to the content of the website with the aid of new tools, examples of these being social media and blogs.

It should be pointed out here that most of the literature discussing Web 2.0 in the context of digital preservation comes from the archival field. However, it should be kept in mind that

institutions outside of the archive sector who hold digitally preserved material have also taken advantage of this development, mainly in the form of using social media platforms to promote their collections and other public programmes. Another example of the use of Web 2.0 is the utilisation of user generated content through citizen science and other crowdsourcing projects.

It did not take long for the new Web 2.0 tools to be adapted in the archival setting, and by the end of 2000s the first articles mentioning Web 2.0 in the archival sector were published (Pearce-Moses 2007; Evans 2007; Jimerson 2007) (Pearce-Moses 2007; Evans 2007; Jimerson 2007). Palmer (2009) saw its potential, as users were now able to contribute to the archive, engage with it, and play a central role in defining its meaning (Palmer 2009). It became possible to create and exchange user-generated content (Kaplan and Haenlein 2010, 61) and Murambiwa and Ngulube (2011) perceive a shift towards transparency, openness and the liberalisation of access to information held by publicly funded institutions (Murambiwa and Ngulube 2011, 89). Palmer and Stevenson perceived a similar shift and characterised it as follows: *'Web 2.0 has forced archives to open up and not only tailor towards the scholar, but also new and unfamiliar audiences'* (Palmer and Stevenson 2011, 7) and Yakel (2011) even refered to this shift as the *'Second Great Opening'*, the first being when the first critique on the archivist's role happened in the 1970's (Yakel 2011), this will be further discussed in Chapter 4.

This new set of tools and this shift was perceived as leading to a more open and inclusive archive, which was called Archives 2.0, a term coined by Theimer (2011):

> *'Archives 2.0 is an approach to archival practice that promotes openness and flexibility. It argues that archivists must be user-centred to embrace opportunities to use technology to share collections, interact with users, and improve internal efficiency. Archives 2.0 thinking incorporates measurements and assessment as essential tools and bases procedures on established professional standards and practice. It requires that archivists be active in their communities rather than passive, engaged with the interpretation of their collections rather than neutral custodians and serve as effective advocates for their archival program and their profession' (Theimer 2011, 59).*

Theimer (2011) reflected on a changing attitude towards being more user-centred. Pearce-Morris (2007) had already pointed out that Web 2.0 was not only about developing new skills

and knowledge, but also about changing one's attitude (Pearce-Moses 2007, 19). Abram (2007), Jimerson (2007) and Evans (2007) had mentioned similar ideas, stating that archives should reinvent themselves to make this work (Abram 2007, 164–66; Evans 2007, 400; Jimerson 2007, 281). Palmer (2009) stated that Web 2.0 tools highlighted a problematic dichotomy between 'user' vs. 'archival authority' (Palmer 2009) and Gerencser (2011) argued that by using these tools, *'…archivists are building new audiences and user communities while raising the visibility of their collections and themselves'* (Gerencser 2011, 177). The main theme emerging in this discussion then seemed to be, that to fully embrace these techniques, the attitude of the archivist had to change (Townsend 2011, 213). A similar point was made by Baxter and again highlighted by Jimerson (Baxter 2011, 299–300; Jimerson 2011, 213).

The response to Web 2.0 by digital preservation practitioners, which are mainly archivists in this case, can therefore be seen as slightly different to their response to Web 1.0, in that it seems to have been more of an intellectual one than a technological one. Instead of just trying to cope with the technology, they saw Web 2.0 as a prompt towards a more active and inclusive approach to archiving as these new tools offered an easy and low-cost solution to opening up to new audiences and communities.

However, Web 2.0 tools have only been around for a decade, and Liew et al. (2015) considered that the use of these tools is still in the 'experimental phase' (Liew, King, and Oliver 2015). In the last decade, a number of surveys have been conducted concerning the adaptation of Web 2.0 tools in the archive sector. A survey by Duff et al. in 2013 pointed out that some participants felt that the ability of users to contribute knowledge would have a democratising effect on archives as a broader perspective of an event would be documented. This comment is in line with the concept of Archives 2.0 and what the sector would like to achieve by implementing these tools. Additionally, the survey showcased that the main use of social media by archives was to promote events and resources (Duff, Johnson, and Cherry 2013, 89).

Another survey by Liew in 2016 explored how and why institutions engaged with social metadata, and concluded that the three main reasons were: increasing public engagement, improving discoverability of items/collections and understanding public interests (Liew 2016, 125) (Liew 2016, 125). Nevertheless, in most cases, this social metadata or user-generated content is still treated as supplementary and kept separate from the content created by the archivists themselves (Duff, Johnson, and Cherry 2013, 80). From the numerous studies

conducted by Liew and colleagues it is concluded that user-generated content is not deemed as something that should be preserved or important (Liew, King, and Oliver 2015, 8). As Liew (2016) says: *'…it does not necessarily increase the inclusivist and many archives are not even aware of this and is not their main motivation of doing it'* (Liew 2016, 133).

Concerns are therefore raised on how much Web 2.0 tools has helped change the attitude of digital preservation practitioners in respect of their responsibility for access. The 2.0 concept, which mainly results in social media engagement in practice, does seem to help break down barriers, but it is unclear how much this approach is providing improved access to digital material in the online space.

3.3 Web 3.0

Integration has become a popular term following the emergence of Web 3.0 (also known as the Semantic Web) and Linked Data. These terms were coined by Berners-Lee et al. (2001) who envisioned a virtual web, called the Semantic Web, where everything was connected through Linked Data (Berners-Lee, Hendler, and Lassila 2001). This is of specific interest to repositories who may have worked in isolation from each other and have unique collections. With this new vision repositories can imagine a future where different collections and databases can be linked up to make it possible to address broader research questions (Netwerk Digitaal Erfgoed 2019; The National Archives 2017b; Towards A National Collection 2021). Larger platforms such as Europeana and ARIADNE are integrating datasets across Europe (Europeana Foundation 2022; ARIADNE 2017).

Currently, the environment in repositories is database-centric and traditional knowledge organisation strategies are used, such as metadata schemas, thesauri and taxonomies (Li and Sugimoto 2018, 37–38). These traditional strategies force repositories to pick from the available options to describe data within the context of an established field with an agreed viewpoint, making it very limiting for repositories in what they can describe. Furthermore, this makes it difficult to integrate datasets with each other, especially when they use different metadata standards with different descriptions and viewpoints. As Bruseker et al. (2017) argued, integrating these datasets from different repositories within larger subdomains or the whole cultural heritage sector is nearly extremely challenging when using a database-centric approach (Bruseker, Carboni, and Guillem 2017, 108–10).

Linked Data holds the potential to solve this problem. It offers a different approach to the data with a high level of abstraction at the top. This level of abstraction is built upon fields of specialisation underneath these higher levels, which can be expanded slowly to include

subdomains. An example of this would be the CIDOC Conceptual Reference Model (CRM) which is a tool to enhance information integration within cultural heritage settings (Bruseker, Carboni, and Guillem 2017; ARIADNE 2017).

Digital preservation practitioners are aware of the opportunities that Linked Data and the Semantic Web have to offer and have started to experiment with using these methods in practice. Niu (2016) surveyed the implementation of Linked Data within archives. She concluded that for the most part, Linked Data is still in the early stage of implementation. However, most archival institutions were and continue to convert existing descriptions rather than producing original ones. Some data models for archival Linked Data are created based on existing archival description standards, such as ISAD(G), which also has a Web Ontology Language (OWL) version, making it possible to be adapted for Linked Data (Niu 2016). The OWL extension of ISAD(G) seems very similar to Records in Context (RiC) but has had negative feedback based on similar issues raised by Niu where the focus lies on converting existing description and practices (The National Archives 2016). Furthermore, re-using controlled vocabularies in a Linked Data context requires substantial knowledge of why and how they were constructed and how they evolved across the decades (van Hooland and Verborgh 2014, 135).

Atom 3 (AIM 25 et al. 2018) and the Person Name Vocabulary (Petram, Dechesne, and Kruithof 2018), are two examples of controlled vocabularies used within the digital preservation community that are aware of the more fundamental shift of thinking that is needed to make Linked Data and therefore the vision of the Semantic Web work. A number of digital repositories have experimented with implementing Linked Data in practice, including DANS in the Netherlands (Meroño-Peñuela et al. 2018) and the legislation service run by TNA (The National Archives 2017c). Hawkins (2021) also saw the benefits of implementing this technology to increase access to material (Hawkins 2021).

The Web 3.0 evolution of the Web is characterised by a view of data as something that is a crucial part of the infrastructure. Currently digital preservation practitioners are struggling to make this a reality and the use of Linked Data has not been adopted to its full potential.

3.4 The Digital Shift

Currently there is frustration felt around how digital archival material is made accessible. This is touched upon when discussing the critique of Web 2.0 and Web 3.0, where these technologies have not been fully adapted because of the attitude of the digital preservation practitioners towards the material. In the light of the implementation of these technologies,

the push for a digital shift has been discussed by numerous researchers within the archive and library sectors. Cook was one of the first to raise this idea, writing that: *'We have paper minds trying to cope with electronic realities'* (Cook 1994, 302). Theimer (2018) called for archivists to become *'masters of data'* (Theimer 2018) and Moss et al. (2018) made a similar call, asking for a change of perspective in repositories to start seeing their material as data to be mined (Moss, Thomas, and Gollins 2018, 120). Then again similar ideas are expressed by Mordell (2019) and Bourg (2017), who echoes this and calls for a different way of thinking about digital material (Mordell 2019; Bourg 2017). This new way of thinking included seeing the digital archival material, or data, as an entity in itself, which needed a different way of thinking from the paper components that many practitioners were familiar with.

Since the early 2010s institutions are showing a growing interest in tackling this digital shift, which is seen for example in the growing number of labs. Labs are dedicated spaces, mainly within larger organisations, which open up the possibility to experiment with digital material in different settings (Mahey et al. 2019). TNA sees this digital shift as one of the biggest challenges facing the archive sector (Goudarouli, Sexton, and Sheridan 2018). The volume of digital material with which this sector needs to deal is of concern to individuals looking after digital material, and some see AI as the only solution to being able to process and provide access to it (Colavizza et al. 2021). Coleman (2017) saw the potential of computational methods and how libraries should be adapting (Coleman 2017) and a number of reports and articles proposed the introduction of computational thinking into archival education or the humanities as a whole (Underwood et al. 2018; Marciano et al. 2018; Colavizza et al. 2021). This will hopefully ensure that the next generation of humanities scholars, and therefore most digital preservation practitioners of the future, will be able to engage more with this material.

For practitioners currently working in the field, there are already a number of courses that are available, such as the Postgraduate Certificate in Applied Data Science at Birkbeck (Birkbeck 2022). Additionally, a number of online sources are available. The Library Carpentry offers lessons and workshops to improve the data skills for people working within library and other information-related disciplines (Library Carpentry 2022). The GLAM Workbench, a collection of examples to highlight opportunities of digital material (Sherratt 2020), and The Programming Historian, a peer-reviewed academic journal for digital humanities and digital history (The Programming Historian 2020) are more tailored towards users, but do give digital preservation practitioners an idea of what may be possible with their material. This different way of thinking about the material, and the new possibilities that the development

of technology offers has led to some changes in practice. Datasets have become available in bulk and Application Programming Interfaces (APIs) have been implemented in many places (Edmond and Garnett 2015; Tasovac et al. 2016).

The movement towards providing access to digital material as data did not however happen in a linear fashion and the degree to which it has happened has depended on a number of factors including the current infrastructure of the organisation, the technical possibilities of the time, the resources in terms of staffing available to the organisations and the nature of the material. It should be pointed out here that this change towards making material available as data can be seen in repositories that have a longer history of providing access to datasets, such as the UK Data Service (UK Data Service 2022). Even in these cases, there has still been a more traditional form of access with datasets being listed in an online catalogue and then accessed and downloaded individually. Again, just as with the development of the online presence of digital repositories and the history of digitisation, there is a focus on the future and new technologies meaning that older implementations are badly documented or not documented at all.

Reviewing the literature, it becomes clear that a different way of processing and accessing digital material is starting to emerge. Even though digital material has been around for a few decades now, and even though the use of quantitative methods has been explored before by humanities scholars (Reynolds 1998), this is the first time that technology has made it possible to achieve processing and accessing of digital material on such a large scale. The necessary technology is far more scalable and affordable with developments such as cloud computing (Amazon 2020), the introduction of Jupyter Notebooks (Project Jupyter 2020) and GitHub repositories (GitHub 2020) to name a few examples. All these technologies make it far easier to process and provide access to digital material online in bulk. The section below will go into more detail of the new approach this has facilitated.

3.5 Computational Access

Making material available as data is still novel but is starting to be explored in the digital preservation community. For example, a different way of exploring material is proposed by the 'Collections as Data' project. The final project report outlined different institutions that have been making their material accessible in a *'computationally-driven'* way, but also gave guidelines to institutions that want to start providing this type of access but are not sure where to start (Padilla et al. 2018). This project had a strong focus on making material available in a computationally-driven way, but did not discuss how this approach can be

accomplished differently depending on the type of material the institutions are making available, or the capacity or resources that institutions may need to enable this type of access.

Another example is the grant programme 'Digging Into Data' which ran between 2009-2020 and considered new ways of exploring big data for the humanities and social sciences, with a focus on the research community. The project outlined a broad range of institutions interested in working on material at scale (Digging into Data Challenge and Trans-Atlantic Platform 2019). Both 'Digging into Data' and 'Collections as data' have an emphasis on state-of-the-art algorithms or projects that are more naturally thought of when thinking about computational methods, such as computer vision, text mining or text analysis. However, a lot more seems to be involved when examining the definition of 'computational':

> *'Involving the calculation of answers, amounts, results, etc' (Cambridge Dictionary 2020).*

This definition only refers to using some type of calculation as part of the work; it does not even imply the use of a computer or a calculator, even though this may make the task easier. This suggests that computational access to digital material could have a much broader range than is sometimes considered, e.g. from users creating visualisations of data, to calculations of averages within a spreadsheet or even by hand, to the more sophisticated machine learning techniques which are popular today. However, what all these methods do have in common is that access to a larger amount of material is needed to achieve them. In this research therefore, the term 'computational access' is not taken to imply only the use of *state-of-the-art algorithms*. This is felt to be exclusionary in the same way as D'Ignazio and Klein (2020) consider the current definition of data science to be exclusionary by leading to an emphasis on *'…formal credentials, professional affiliation, size of data, complexity of technical methods, or other external markers of expertise'* (D'Ignazio and Klein 2020, 14). Currently there is a preference within the digital preservation community towards novel computational methods, but computational access should not be limited to funded researchers with a novel research project but should instead be available for any user who may want to access material in this way.

Computational access, as defined above, is still in its infancy within the digital preservation community, but there is a growing interest in it across institutions. Bailey (2018) touched upon different models that institutions may employ to make such access possible, but again with a strong focus on research and institutions. The underlying assumption seems to be that

academic researchers have funding or are part of a project which makes it possible to have a bespoke access tool created for them, or for a developer to be employed to help them; this is not a feasible option on a larger scale. However, there is still an assumption that some of this access can be fulfilled physically, by exchanging hard drives in person for example (Bailey 2018), which is again not feasible when making computational access more widely available.

Based on the above review of the current thinking about and application of computational access, this thesis proposes the following categorisation of approaches; datasets, Application Programming Interfaces (APIs) and platforms. This thinking around computational access and the proposed categories were developed as a result of the collaborations undertaken during the PhD, e.g. work with Mark Bell on the Machine Learning Club (Bell and Talboom 2022) which was extended during my SSI Fellowship and resulted in the publication of the Computational Access Guide (Talboom and Digital Preservation Coalition 2022). Below a short summary of this work can be found, as this context is of interest to this overview.

With the dataset approach, datasets are created by institutions and made available to users who can then take them away to compute over in their own environments. The datasets are normally made available in either a Comma Separated Value (CSV) or JSON format, as both are commonly understood and easy to read by both humans and computers (Drapeau 2018). This approach to computational access gives the organisation control over the material being made available, but it does need maintenance as the datasets have to be updated and reloaded. A few examples of such datasets are those made available by the Museum of Modern Art (MoMA) (MoMA 2020), the Carnegie Museum of Art (CMoA) (CMoA 2017), and the extracted feature datasets of HathiTrust (HathiTrust 2020).

The API approach makes it possible for users to send a list of instructions to a data store, which is usually a server and a database maintained by the organisation holding the digital material. This list of instructions is then processed and data are returned to the user (Hoffman 2018); therefore creating a more fluid way to provide access to data than the datasets approach, as the user can be more specific on the data that they want and an API requires less bandwidth and disk space. Examples of this approach are the APIs run by The National Archives (Underdown 2018; The National Archives 2021a) and the Wellcome Trust (Wellcome Trust 2020), but also APIs such as Europeana's search API (Europeana Foundation 2020).

The last approach, which is through a platform, is mainly reserved for material affected by copyright or other legal restrictions. The platform normally takes the form of a dedicated

online environment where the user is able to run the required computation. The infrastructure behind this approach can be quite similar to the other two, but the main difference here is that not only does the user have access to the data, but also access to a set of tools to run over the data. Examples of the platform approach are the CLARIAH media suite (Melgar-Estrada et al. 2019; CLARIAH 2020) and the Archives Unleashed project (Ruest et al. 2020). The control around these environments varies; some only provide access if you are a registered user whereas others are open to anyone. Additionally, it may be possible to manipulate material in the environment, but not necessarily output any results; this could be because of copyright or other legislative restrictions.

At the moment notebooks seem to offer repositories one possibility for providing access to their collections as data. Notebooks are web-based platforms which can be used to run code in the browser, the most popular options being Google Colab (Google 2020) and Jupyter notebooks (Project Jupyter 2020). The main difference between these two applications is that Google Colab is cloud-based whilst Jupyter notebooks is not. This means that users do not have to download or install anything locally when using Google Colab, whilst this is not the case for Jupyter notebooks.

Examples of notebooks within the community are the GLAM Workbench (Sherratt 2020) and The National Library of Scotland, who have recently released a set of notebooks that allow users to explore their collections (National Library of Scotland 2020; Ames and Havens 2021). Notebooks have grown in popularity due to being an ideal environment for experimenting with different techniques; furthermore they are easily accessed and sustainable enough for their purpose (Candela et al. 2020). Whitelaw (2015) argued that this is the way forward and names this approach 'generous interfaces' (Whitelaw 2015). The CLARIAH project gave an example of using Jupyter notebooks to showcase a generous interface similar to Whitelaw's proposal (Wigham, Estrada, and Ordelman 2019).

Computational access is a new way to provide access to digital material which is being explored by the digital preservation community and which, whilst currently still very much in the early stages, is starting to show promise. It currently has a very academic focus, which may make it difficult to translate across the whole digital preservation community.

3.6 Perspectives to envisioning access

The above sections have set out the contexts against which the work in respect of providing access to digital material of the digital preservation practitioners interviewed takes place. One of these contexts is the evolution of the Web. Web 1.0 enabled some type of presence

online; Web 2.0 invoked a more intellectual response and Web 3.0 emphasised the importance of thinking about digital material in a more computational way. The current state of the art in respect of computational access provides another relevant context and was set out in the above section.

The following section will report findings from the interviews which focus on how the participants are envisioning access in their current institutions as well as how they would like to envision it in the future. The main theme discussed in this section is around the digital infrastructure and will focus on the following sub-themes: 'a sustainable digital infrastructure', 'collaboration', 'not only technical' and 'envisioning a solution'. Headings will be used within this part to reflect these sub-themes.

As discussed in Chapter 2, participants from a broad range of different institutions were interviewed, from smaller organisations including community and specialist archives to larger national institutions, such as TNA and the British Library. The infrastructure of an organisation can be both a benefit and a drawback. Depending on the specific organisation there could be more room for experimentation and less legislation to keep in mind. Then again, larger organisations may have to deal with more bureaucracy in general, but as a counterpoint are likely to have access to more resources and skills that could support access to digital material.

Below findings are discussed outlining the main themes that participants brought up in respect to discussing the infrastructures of their employing organisations as well as some proposed ideas and solutions that they have envisioned in terms of their specific situations. The participants whose perspectives are reported in this section are from Study Group 1 and Study Group 2, as both these groups work within the digital preservation community.

3.6.1 A sustainable digital infrastructure

Firstly, with regards to the theme of existing infrastructures, concerns were expressed mostly by participants from Study Group 2 with more technical backgrounds. One of the first things they critiqued was the current way that digital infrastructures and projects are funded in many organisations. Ordelman from the CLARIAH project is very honest about his previous position at the Sound & Vision Labs, where he found the projects being undertaken to be outdated and very niche. Furthermore, he talks about how hard it is to re-use the tools developed within the Labs, as most of them have been built for very specific purposes. Tweed also highlights this issue:

> *'What you are basically saying, if you've got Labs and you've got all of your best people working in a little innovation team called Labs, what you're saying is: our institution is broken because we can't actually improve our actual stuff. But that doesn't sound as good as like: our Labs team made a thing. Well, that's great, but what about your actual site that people actual use that hasn't changed in fifteen years?'*

Ordelman, from the CLARIAH project, has similar concerns around Labs. He hopes to learn from past mistakes and create a far more sustainable platform within the CLARIAH project that will facilitate the use of digital material for a large range of purposes. Ordelman hopes to showcase how such projects should be done by drawing in more technical expertise. He questions the absence of technical skills on some digital research projects: *'Huh, that's strange, the number of researchers involved is far larger than the number of developers that are working on the project. And that's strange because we are developing an infrastructure, which is technology and you need a lot of programmers for doing that work right?'*

In the same vein, the Wellcome Trust have recently moved away from having many bespoke websites to access their digital material and are now working with one underlying infrastructure on which a number of interfaces are built. Tweed thinks this is a far better solution, partly because it ensures that the core technology needed is funded through core funding and not through small research projects:

> *'As somebody's research, as somebody's project, that's kind of okay. But it's when institutions think that they can use that model to do their long-term looking after their stuff, that's not so okay. And there's reasons why successful tech companies, companies in tech do not work like this.'*

He emphasises the importance here of seeing this infrastructure as the core of the institution to ensure longevity is guaranteed for the technology. And finally, the British Library has also stepped away from using bespoke projects and have been developing core systems and workflows, as Pledge points out: ''*I would say firstly, we don't really work with projects. One of the problems with projects is that they tend to rely on, in my experience, funding. There's a short-term burst of work and then the project ends or tails off, but often with things unfinished or incomplete.'*

3.6.2 Collaboration

In the context of this concentration on creating more sustainable infrastructure, another theme emerged around the importance of working collaboratively with other organisations.

The main concern here seemed to be that not every organisation would have the skills to create a digital infrastructure and that other organisations could potentially help with that. For example, Tweed spoke as follows: *'Right, and then it just comes down to resource and where that can be best spent. And I just don't think that is best spent within individual institutions trying to solve these problems on their own.'* Ordelman is hoping to offer a solution to smaller organisations, if possible, with the CLARIAH project: *'But the global idea is exactly what you say that you don't want all these smaller institutes making a complete mess out of it and that we should centrally organise it, so that we can also provide trust, authority, on both the data level, but also on the technology level.'*

Furthermore, there is a concern with creating tools that are beneficial to a wider range of people. CLARIAH is doing this by experimenting with new user groups. But also, Milligan from the Archives Unleashed project, is able to showcase the multipurpose use of tools not only built for the users, but also for digital preservation practitioners themselves. The interviewees from the International Institute for Social History (IISG) also hint towards using tools with a wider benefit when talking about applying AI to their collections, as they would like to provide tools that are not only useful to the archivists working there, but would also be beneficial to the users, as Gillesse points out: *'…modern analysis methods and entity recognition tools, artificial intelligence, using these principles you could make more intelligent and more automatic selections. Or at least support the researcher with this. And this is where you will see that the selections and the research questions are strongly linked with it other (not only within preservation and selection, but also the researchers are interested in using these digital humanities tools) [Translated from Dutch].'*

3.6.3 Beyond the technical

Technology is slowly becoming seen as something important and deserving of core funding; small research projects are beneficial but not for building key parts of the infrastructure. Working together to be able to provide everyone with a basic infrastructure to provide access is seen as quite important. Infrastructural changes concerning the technology are starting to happen, but a larger number of interviewees point out that numerous problems are not necessarily technical, but rather come down to how the organisation is managed. If senior leadership within cultural institutions do not realise the requirement to completely rebuild or restructure an organisation, it becomes difficult for those in more operational roles to actually try and establish this.

Copyright was identified by a number of participants as a major problem. Claeyssens, talking about how the Dutch National Library underestimated the time that would be spent on

copyright issues, explained: *'…that aspect is maybe the thing that was most underestimated by the Dutch National Library when we started doing mass digitisation and in general for our digital collections.'* They now have a number of specialists in place and ensure that they make future agreements not only to make files accessible on an individual level, but also in bulk, for computational access, as this has caused problems in the past. Tweed talks about copyright and thinks that he is lucky working at an organisation that has very strong ideas around openness from senior leadership. Without this he is dubious about what would be possible; *'…it's amazing what you can get done if you have somebody above you supporting your position.'*

Not only copyright, but legislation more widely is also seen as an issue. McKean, at The British Library, has been experimenting with ePADD which is open-source software developed by Stanford University that helps with the processing of email archives (Stanford University's Department of Special Collections & University Archives 2021). This trial run of software by the McKean has highlighted that: *'…the technical solution is there, the legislative solution is often not there.'* The IISG talks about issues not necessarily being technical, but ethical. At the moment they may be unsure what people could take from their collections, especially when it comes down to born-digital material, but Gillesse does think that the responsibility for unearthing material lies with the researcher: *'Other groups, for example the police or forensics, which use different methods, they may unearth stuff from the data we couldn't even imagine. This rises a bunch of new questions, especially when it is about ethical stuff. And if the researcher does this, then the responsibility lies with them, of course. We cannot by solely responsible as preservers of that information [Translated from Dutch].'*

Corti discusses how the technology is not the constraint and that it mainly comes down to ethical issues, especially when considering complicated ethical issues which computers or automation will not necessarily be able to solve:

> *'The problem is, the price is not in the hardware at all, it's in the administrative process to have all these things checked and yes, you could automate more, but it is very hard to automate the checking of a ten page form when it's the content that's got to be read by humans and appraised to make sure it's ethical enough and the ethical assessment is really quite difficult.'*

The perspective of the participants is that copyright and other ethical issues are things that need to be taken into consideration when making material accessible and that the degree to

which these can be successfully negotiated depends in some part on the organisation you work for and how well it is organised. Most of the quotes above come from quite large organisations, but similar concerns were found in the interviews with those who worked as smaller organisations.

Another issue raised by the participants was the question of open-source tools. Many participants saw the use of such tools as a way of being sustainable. For example, the GLAM workbench showcases how Sherratt is able to maintain and update a sophisticated platform, something which would have not been possible if he did not have access to open-source tools.

IISG also see using open-source tools as beneficial and have implemented Dataverse for their Linked Data, which they see as a better option than building a bespoke system. Dataverse is an open piece of software for creating a repository for research data (King 2007; The Dataverse Project 2022). They used to create custom scripts, but this was not scalable, so they switched to Dataverse. They see it as important to use open-source tools, as it adds to the community, as Ruijter explains: *'Not everything has to be open-sourced, but it is something we strive towards. It is something nice to strive for of course, because you contribute to something larger than just you own organisation or to just increase the profits of a company. So yes, if we can, we do it gladly [Translated from Dutch].'*

The examples above are about taking existing open-source tools and implementing/hosting them, but some organisations have been experimenting with putting their data on open-source services themselves, such as WikiData. As Tweed states, the advantage of this approach is that it '*…also opens it up to massively different audiences.'* The Dutch National Library also experimented with WikiCommons, uploading material to offer users a different point of access. However, not every organisation will be able to use these approaches, as material needs to be licence free for Wikidata and WikiCommons.

Although the participants were generally positive about the open-source approach, they did express some concerns around its sustainability. Sherratt in particular spoke of how not all of these tools stay available and can change at short or no notice. The way he is currently dealing with this is to provide several access points trying to ensure that there is always something available to users: *'…having it openly available and available in multiple places, is the best thing I can do in terms of long-term use.'*

Organisational infrastructure depends on a number of things, but it seems that organisations are starting to realise that the provision of access to digital material will need major changes to their core infrastructures, which cannot be funded or sustained on a project basis. Then again, it is widely acknowledged and understood that the provision of access is not just a question of technology, but also a legal and ethical matter. For smaller organisations or projects it may be easier to operate under the radar in respect of these last issues, but different problems are then experienced around a relative lack of resources and skills. The use of open-source tools is seen as both useful, but also problematic, especially in terms of sustainability.

3.6.4 Envisioning future access

During each interview questions were asked about how the participants would envision future access to digital material, or what they would like to see being changed. When discussing this with Study Group 1, the digital preservation practitioners, there seemed to be a consensus that the current way of providing access was not working. Numerous participants mentioned that current practices were not as they would wish, but they also had some difficulties articulating how the situation might be improved. The technical people from Study Group 2, on the other hand, could articulate this very clearly. They saw digital material as data, and they ideally wanted to provide access via an API.

The advantages of such an API-first approach were set out very clearly by Tweed from the Wellcome Trust: *'They are both operating off the same data, the Catalogue API, but they are providing two completely different views that are aimed at slightly different users with slightly different search intents. What we are about to start to bring in is discovery based on concepts and slightly more browse-based discovery. Again, that's the same data, that's just adding additional functionality to the same API.'*

Furthermore, the Dutch National Library, CLARIAH project and TNA's Legislation service all work with an underlying API infrastructure, but the API itself is only directly accessible to people who know what they are working with, mainly developers or other technical people. Those without these skills can still access data from the API but via additional interface(s). The Technical Lead of Project Alpha at TNA sees the API approach as the best one for TNA in the future as it makes it possible to deliver multiple interfaces to the user. He discusses how multiple interfaces for applications are already day-to-day practices for many of the tools that he uses and it should not be different when providing access to material. He especially worries about trying to provide everything through one interface*: 'I think you always end up compromising in some way, most commonly by providing too much.'*

The above quotes highlight how the more technical people are envisioning and seeing a more sustainable infrastructure in the API-first approach, which can be beneficial to repositories as it would help solve a number of constraints around access to digital material.

Ultimately though everything comes down to resources. As Dethmers mentions: *'…it's [the infrastructure of the HathiTrust] not modern, but that's not from lack of willpower and desire, it's much more lack of time and resources really. We are pretty well funded, and we could do a lot of these things and there's definitely the willpower to do a lot of it, but we can only do so many things per development cycle. Right now, it's an interesting archaeological look at technology from the past twelve years.'*

3.7 Building new infrastructures

This chapter provides an overview of how the provision of access has been negotiated by digital preservation practitioners primarily through a technological lens. This lens was articulated against the contexts of both the evolution of the Web and the current emergence of a concern with computational access. An important finding from the exploration outlined in this chapter is that it is no longer possible to consider the provision of access without also considering the provision of an often complex technical infrastructure, as highlighted by some of those interviewed. This perspective was provided most clearly by those with more technical backgrounds.

These individuals considered such infrastructures as being a core part of any service and they were therefore very critical of an approach that saw its development dependent on project-based funding. Rather it was their firm belief that to create a sustainable and durable infrastructure for access, its development needed to be funded as part of the core business of an organisation. Another theme that emerged was an emphasis on working together whenever possible. Through the technological lens, this theme manifested in a discussion of open-source tools. Such tools were preferred in contrast to proprietary software or tools, but there was also caution expressed around the fact that their use involved a dependence on external stakeholders which could cause problems in the long run.

The preferred approach for providing sustainable and interoperable infrastructures for the provision of access was, according to the more technical participants interviewed, an API-first approach. This was seen as the best option, because it allowed for many different uses of the digital material through the building of multiple interfaces on top of the API service.

One final finding from using the technological lens was that there did seem to be some clear differences of perspective apparent within the two study groups of digital preservation

practitioners. Those who were less technologically proficient or engaged expressed more uncertainty and doubt. Although there was a clear consensus that digital material needed to be approached differently, they struggled with articulating what this would look like in practice. This concern is also highlighted in the consulted literature in this chapter. Especially around the need to think about this material in a different manner and the difficulties around adapting newer technologies such as Web 2.0 and Web 3.0. On the other hand, those who were more technologically proficient or engaged were able to give a very clear articulation of what it should look like in terms of the construction of a sustainable and interoperable infrastructure in the form of an API on top of which multiple interfaces were built. The next chapter will explore this difference of perspective in more detail as it considers in more detail the roles and responsibilities variously felt by those digital preservation practitioners who participated in the study.

4 Role of the digital preservation practitioner

The previous chapter focused on the provision of access through a technological lens. An important finding was that it is no longer possible to consider the provision of access without also considering the provision of an often complex technical infrastructure. This is clearly understood by the more technical interviewees, mainly from Study Group 2, who see their role as building such an infrastructure, which should be as sustainable and interoperable as possible. All those interviewed however recognised that even when focusing on technological aspects, other issues such as legislative and ethical constraints cannot be seen as separate from this work.

Some differences of perspective were identified within the digital preservation practitioners interviewed as Study Groups 1 and 2. This difference was seen between those participants who were very focussed on building an infrastructure to provide access and those who perhaps viewed their role in providing access in much broader terms. To deepen understanding of these differences, this chapter will explore how the provision of access is being negotiated through the lens of the roles and responsibilities of digital preservation practitioners, that is to say how such practitioners articulate the provision of access as a task that fits within their role.

To that end, this chapter starts by constructing an overview of the emergence of the digital preservation community with a focus on Europe, Australia and North America (see Chapter 2 for further details on the scope), before bringing in perspectives from interviewees as a counterpoint to this picture.

4.1 The digital preservation community

As mentioned in the introduction, the first need for preserving and making digital material accessible for re-use was established in the 1960s within the social sciences (The Steinmetz Archive 1989; GESIS – Leibniz-Institut für Sozialwissenschaften 2019; ICPSR 2019). From the 1970s this need was increasingly acknowledged across other disciplines and across libraries and other memory institutions (Doorn and Tjalsma 2007). From this time onwards at least two streams could be identified who had an interest in preserving digital material. On the one hand there were those individuals who came from a tradition that was familiar with safeguarding objects for long-term preservation, such as archivists, librarians and museum curators. These individuals were mainly based in GLAM institutions and saw this material in terms of objects. And on the other hand, there were also individuals who came from a long tradition of looking after computer-dependent data, so examples of this are scientific data

centres, but also the social science centres mentioned above. These individuals mainly saw digital material in terms of data (Lee 2010, 4020). But it was not until the 1990s that these two streams came into contact with the first attempts to formalise the practice of digital preservation.

The drive towards formalisation mainly came from the library and archiving field. At the end of 1994 the Commission on Preservation and Access and the Research Libraries Group (RLG) created the Task Force on Archiving of Digital Information. In 1996 this Task Force produced a report titled 'Preserving Digital Information' and outlined the concerns that were felt by repositories outside of traditional archives and libraries, such as the social sciences repositories, preserving digital material. It was suggested that these repositories, claiming to serve an archival function, should be able to prove their reliability, ideally through certification (Garrett and Waters 1996).

The concern of 'others' (not from a library and archive tradition) preserving digital material led to a number of influential articles, partly in response to the Task Force. These articles outlined ideas around best practices and perspectives on what was required to preserve digitals material. Haynes et al. (1997) provide an early example of working out what was needed in practice (Haynes et al. 1997) and Hedstrom (1998) was also concerned about repositories not acknowledging the long-term end goal of making digital material accessible. She described digital preservation as:

> *'…the planning, resource allocation, and application of preservation methods and technologies necessary to ensure that digital information of continuing values remains accessible and useable' (Hedstrom 1998, 190).*

Additionally, the article highlighted how there was a concern regarding the medium and formats that digital material was held on, as it changed much faster than the analogue equivalent, therefore raising concerns around the long-term preservation of this material on fleeting formats (Hedstrom 1998). Finally, Hodge (2000) released an article outlining best practice and reiterating similar ideas to those Haynes et al. (1997) and Hedstrom (1998) proposed (Hodge 2000, 200).

Figuring out good practice seemed important in the field at this time, especially in light of the concern that the digital material was disappearing at much quicker rates than analogue equivalents due to the issue of technological obsolescence. This concern even made some speak in terms of our entry into a 'Digital Dark Age' (Kuny 1998). In the early 2000s a

mechanism for formalisation emerged in the form of a reference model for a so-called Open Archival Information System (OAIS). This model was published by the Consultative Committee for Space Data Systems (CCSDS), who developed a set of recommendations for archiving digital material for long-term preservation (CCSDS 2002).

The 'open' in the OAIS model refers to the open forums that were used to develop the model. These forums started in the mid -1990s, highlighting that it was not only GLAM institutions who were concerned about keeping digital material safe for the long-term. By making the development process of this model open, it was possible for institutions beyond the CCSDS and its immediate stakeholders to contribute and comment on the model, making it easier to adapt across different fields engaging with digital preservation. The model provided a higher conceptual framework for organisations to work with, which was needed at the time (Lee 2010). It provided a common model and vocabulary that could aid communication between all those concerned with preserving digital material, not just those from the library and archiving communities, but also those from the other stream who looked after this material as data.

OAIS can be seen as the initiative that allowed ideas, concerns and concepts from both fields to be combined thereby laying the groundwork for a digital preservation community to converge with individuals from a range of backgrounds brought together under the same umbrella in their use of the model. The OAIS references model was updated in 2012 (CCSDS 2012a), and has also become an ISO standard - ISO 14721 (CCSDS 2012c). A third version is currently being drafted (CCSDS 2019). For over two decades now this model has been used by digital repositories and has become a staple when talking about digital preservation.

OAIS has been widely adapted and built on by different institutions, certifications and protocols (Lavoie 2014, 2). It has had many positive effects, including the creation of a shared terminology; the majority of digital preservation practitioners will be familiar with terms such as the 'Designated Community' and the 'Archival Information Package'. Furthermore, OAIS has had influence on many developments, including topics such as file format checks, persistent identifiers and fixity checks (Sierman 2019, 39–44).

Around the same time as the development of the first version of the OAIS model, a joint report from the RLG and the Online Computer Library Center (OCLC) regarding the development of a certification program and auditing process titled 'Trusted Digital Repositories: Attributes and Responsibilities' was published (RLG and OCLC 2002). After a few years of discussion an updated version of this report was released which was titled

'Trustworthy Repositories Audit and Certification: Criteria and Checklist', better known as TRAC (CRL and OCLC 2007). This was later revised and updated in 2011 and in 2012 TRAC was incorporated in the ISO standard 16363 (CCSDS 2012b).

In the late 2000s it was apparent that people from many different backgrounds could identify as a digital preservation practitioner. Furthermore, the more traditional fields became more open to this understanding of them not being the only ones capable of looking after digital material. This is especially seen when looking at definitions from within the field. In 2007, a new definition of digital preservation offered by the American Library Association (ALA) highlights different needs and acknowledged that digital preservation is not only done by libraries and archives, but also by allied professions (American Library Association 2008). Then again, the Digital Preservation Coalition (DPC), which has been around since 2002, has a growing number of members from all different types of backgrounds and defined a more inclusive term for digital preservation, mainly focusing on the technical aspects of things: *'The process of storage, backup and ongoing maintenance as opposed to strategies for long-term digital preservation.'* (Digital Preservation Coalition 2015). This definition is from their handbook, which gets regularly updated.

Certifications and standards keep playing a large role within the digital preservation community. Besides TRAC, there are a number of other auditing tools that are noteworthy. The Data Seal of Approval, which was created by the Data Archiving Networked Services (DANS) archive in the Netherlands, has recently been updated to Core Trust Seal (Core Trust Seal 2021). Additionally, there is DIN 31644 in Germany, better known as the Network of Expertise in Long-Term Storage of Digital Resources (NESTOR) seal (Nestor 2019). The Digital Repository Audit Method Based On Risk Assessment (DRAMBORA) model (Donnelly et al. 2009) and the Simple Property-Oriented Threat (SPOT) model (Vermaaten, Lavoie, and Caplan 2012) are two other examples of models that can be used to audit a trustworthy system. With 16 guidelines, the Core Trust Seal has the least amount of guidelines, whereas ISO 16363 is the most demanding with 109 guidelines.

The field of digital preservation is made up of actors from a range of different disciplines. Early on the streams of those with an interest in digital material could be seen as more distinct, but they have been drawn increasingly closer. One of these streams was the GLAM sector which has a long history of preserving material. The other consisted of actors that had held digital material from the late 1960s and saw it more as data. A number of models and certifications, and in particular the OAIS model, were introduced to be able to communicate

between these different streams and it is now widely acknowledged that digital preservation practitioners can come from a range of backgrounds, the one thing that they have in common being that they preserve and look after digital material for the long-term. This is evident when consulting the member list of the DPC, which predominantly consists of libraries and archives, but also institutions dealing with mainly data are present such as the UK Data Archive and the Atomic Weapons Establishment (Digital Preservation Coalition 2022a).

4.2 Current constraints due to the OAIS model

The OAIS model has been seen as a way for the different streams of digital preservation practitioners to communicate with each other and therefore build common ground and a common practice around digital material (Lee 2010). However, in the last few years the OAIS model has been receiving a large amount of critique, and this section will highlight that critique. The constraints that will be listed below showcase tensions between the ideas and perspectives within the model and what is happening in practice. It should be mentioned that all the certifications and auditing processes discussed in the previous section have been based on the OAIS model and these certifications and auditing processes have been popular. As of December 2021, two repositories have been formally accredited under ISO 16363 (Primary Trustworthy Digital Repository Authorisation Body 2021), six under TRAC (Center for Research Libraries 2021) and 131 hold the Core Trust Seal or the earlier Data Seal of Approval (Core Trust Seal 2021). The popularity of these schemes can be explained by Sierman and Ras (2019) who point out that archives gain these certifications and go through the accompanying auditing process as they can aid a repository in becoming an organised institution following a set of guidelines and they ensure a centralised view on what a good digital repository should look like (Sierman and Ras 2019, 72).

However, auditing processes and the underlying OAIS model have been receiving critique over the last couple of years. Starting with the auditing processes themselves. These are completed by institutions to be viewed as a trustworthy organisation, but in many cases, it can be unclear what this actually means. McGovern (2016) pointed out that certified repositories view themselves as being trustworthy (McGovern 2016, 329). Yoon (2014) conducted a study into the perception of trust from a user point of view, interviewing 19 users and one of the important findings to come out of this study was that repositories should make it apparent to users what their missions and function are, as this would decrease the uncertainty of users, which can positively influence their trust. There is no mention of auditing processes or certification having an impact on the trust of the users (Yoon 2014). Bak (2016) extended this argument on being a trusted repository:

> *'Though conceptually foundational to the idea of a TDR, the question of whether cultural institutions like national archives are fully trusted by the public is not addressed in the TDR report – it is asserted as a simple fact' (Bak 2016, 380).*

He further critiqued the original RLG/OCLC report from 2002 as being too binary and not addressing trust in archives, calling it a technocratic approach. Following on from that, Bak (2016) mentioned that the report has a strong emphasis on the internal processes of an archive and does not integrate the relationship of trust with external users or partners (Bak 2016, 382). This is emphasised in the report of Sierman and Ras who mentioned that repositories conduct auditing and certification for themselves, not necessarily for the public (Sierman and Ras 2019). But if this is the case, what constraints are the OAIS model and the resulting auditing processes and certification having on the archiving sector and therefore the access of digital material?

4.2.1 Linear Process

The OAIS reference model works on the assumption that digital preservation is a linear process, proceeding from ingest, to preservation, to dissemination. However, as van Essen (2019) pointed out, this is not necessarily the case for digital material. Digital material, and especially born-digital material, is more complex than analogue material, and the boundaries of each separate entity are not as clear and there are far more relationships between files. On top of that, preservation strategies can change over the years; there is not one solution to preserving digital material for the long-term, which means that preservation copies may change (van Essen 2019, 138–39). The dynamic nature of digital material is currently not reflected by the OAIS model, which is raising concerns about its usefulness.

In a talk given by John Sheridan of The National Archives, Sheridan mentioned that the linear process of the OAIS model has another downfall on top of the inability to accommodate the dynamic nature of digital material (Sheridan 2019). Archival practice itself is also changing as given that the provision of access requires the provision of a complex technological infrastructure (as was discussed in Chapter 3) archives are starting to outsource this. One example of outsourcing would be data storage. The UK Government Web Archive (UKGWA) at TNA, the Wellcome Trust and the Archaeology Data Service (ADS) are all examples of institutions where the data storage has been outsourced to a third party, for example Google Glacier. Other parts of the infrastructure can be outsourced as well, for example to Preservica (Preservica 2018) or other digital preservation platforms. However, the OAIS model assumes that the archive is the infrastructure and that all the work of

preserving digital material is carried out by one isolated system; this is not the case anymore for many repositories. The model assumes a perfect system but the current technological ecosystem favours a distributed and outsourced one.

In this context Sheridan (2019) introduced a model developed by Simon Wardley called the 'Product Evolution Cycle'. This model showcases the lifecycle of a successful product, it starts with its genesis, and if it is successful, it evolves, others copy the product and create new custom solutions from it. If the product is still successful, it diffuses further, and others create new products which get improved, extended, and becomes widespread and available, *'ubiquitous, well understood and more of a commodity'* (Wardley 2017, 17). An example of this, as mentioned before, would be the outsourcing of digital storage; why should repositories build a custom-built version of a storage infrastructure when it is possible to outsource storage to a company who offers a better and more robust system? Digital repositories are doing this, but within the OAIS model there is no room for outsourcing. It is not about building robust systems anymore to archive digital data, but it is about trusting others and taking the risk of trusting others with the material or product, the OAIS model is currently not accustomed to this type of system.

The following point will tie into the next argument about the *Designated Community*. As the OAIS model sees the flow of digital material from dissemination to consumer as a one-way process; the consumer can order a *Dissemination Information Package (DIP)* or request documents from the archive and the archive is able to respond to this request. This process does not leave any room for any feedback from users and sees the users as essentially outside the system and thereby unable to influence the knowledge held within the digital repository itself (CCSDS 2012c, 62).

Users hold a vast body of knowledge when it comes to the records held in archives. Kärberg and Saarevet (2016) referred to this knowledge as *User Knowledge*, however, the OAIS model does not give any acknowledgement of this type of knowledge (Kärberg and Saarevet 2016). In practice, many archives offer their users the opportunity to engage with the catalogue by adding tags and suggesting corrections. Many crowdsourcing projects have been introduced, where the user is able to help the archive in extracting knowledge from documents. Examples of this are Transkribus (Muehlberger et al. 2019) or the British Library Flickr account (British Library 2021a), which was created to improve tags for images.

4.2.2 Designated Communities

Within the OAIS model the users who will be using the material after the *Dissemination Stage* are defined as the *Designated Community*:

> *'An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the Archives and this definition may change over time' (CCSDS 2012a, 22).*

For a repository to become *OAIS compliant*, it is considered necessary for them to define a scope for their primary *Designated Community*. The reference model aids in this by giving a definition and emphasising the importance of defining one but does not give any guidance into how to create one in practice. The only thing the model is clear on is that not *'everybody'* may be defined as the *Designated Community*, as that would imply the collection of an unfeasible amount of *Representation Information*, or place an unreasonable expectation on the level of *Knowledge Base* required from such a broad community in order for them to independently understand the preserved digital information (Bettivia 2016, 41).

This stance however causes tensions when a repository has a mandate to serve the general public, and therefore *'everybody'*. If a repository was to define a *Designated Community* that would abide by the rules of OAIS and therefore not include everybody, it would make the data accessible to the world, but only usable by a small number of people, resulting in publicly accessible data, but not necessarily publicly usable data (Bettivia 2016, 42).

Furthermore, Giaretta (2011) explained that if a repository was to only make data accessible to a limited range of people or accessible to the wrong user group, the depositor would not entrust their digital objects to the archives and go elsewhere. Giaretta argued that this keeps a digital repository honest and therefore that they will define the *Designated Community* in a way that will suit their repository in the best of ways (Giaretta 2011, 51). Nonetheless, this argument does not necessarily work in practice; many national institutions have depositors who have no choice but to deposit with them, making it impossible for the depositor to go elsewhere and therefore defining a *Designated Community* may not cater towards the right audience.

Another issue is that digital data is complex, and it can be unclear what the *Significant Properties* of it are. The OAIS model helps in determining the *Significant Properties* by gearing these towards the *Designated Community*, but *Significant Properties* are situational as this

will depend on the audience (Bettivia 2016, 40). Take for example video games; it becomes increasingly difficult to extend access to the general public (McDonough 2012), as users are interested in different properties. Some users will want to use the code, while others would want the actual game experience. Newman (2018) has identified a third potential use for video games with his 'Game Inspector' tool, which looks at gameplay preservation; this looks at preserving different communities or people playing the a specific video game (Newman 2018). The example given focuses on video games, but is applicable to any type of processable data, including 3D models and GIS maps (Locher 2016; 2019), there will be different ways to work with the data, and therefore different ways to make this material accessible and usable.

Related to the discussion around the *Significant Properties* is the discussion of what metadata should be kept with these files to ensure provenance and authenticity, and to ensure the *Designated Community* is able to use the files in their intended way. Born-digital metadata to aid in the contextualisation of material has many more types than before; some state three different types, such as NISO and Miller, but Getty introduced five different categories (Corrado and Jaffe 2014, 36). TNA has recently identified seven different metadata types for born-digital material (Hillyard 2018), which are currently not reflected in their Discovery catalogue. What these categories highlight is the complexity of this material, and the difficulty of capturing the *Significant Properties*.

As well as assuming that the archive is the infrastructure and a system operating in perfect isolation, the OAIS model also seems to assume that it is possible to designate in advance both the user group and hence the use to which the digital material being preserved will be put. To be sure this does allow an OAIS to draw its scope quite narrowly, making it easier to fulfil, but many repositories in practice do not and cannot operate on this assumption.

4.2.3 'Business as Usual'

Adapting the OAIS model has become *'business as usual'*, meaning that institutions have adapted the model and have made it part of their day-to-day practice, where different *Delivery Packages* are created, and a *Designated Community* has been defined. Whilst it is being welcomed that the OAIS model has been so influential and has helped to establish the practice of digital preservation, some critics are now starting to warn that the model is becoming such a normal part of the digital preservation process, that there is little room left for critical thinking (van Essen 2019, 140–41).

Furthermore, little room is left for innovation and experimentation, with a tendency for archives to just tick the boxes asked of them. The OAIS model does not give any clear guidance and is meant more as a conceptual model with very high-level terms explaining the concepts (Sierman 2019, 43–44). This has led to many archives copying certain parts from each other and having very similar descriptions to each other (Talboom and Underdown 2019). The OAIS model was set up as a model for repositories to engage and improve their digital preservation system, however as it is so highly influential, it has led to archives putting in minimal thought when becoming *OAIS compliant*, which is missing the point of why the reference model was developed in the first place.

This lack of engagement with the actual model, and the lack of examples given by the OAIS model has led to little to no innovation or critique on the model over the last few years, and the current draft for Version 3 sees little changes to its predecessor (CCSDS 2019). Digital preservation practitioners should not forget that the model is a reference model and not a ready-made solution to digital preservation, with the main purpose to be able to facilitate as many implementations as possible (Sierman 2019, 38).

The certifications that are based on the OAIS model showcase the value the institutions place on being trusted, but do not necessarily improve the possibilities for access or the trust that users feel towards the institutions (Yoon 2014). Certification seems to provide a way for organisation to feel more trusted, as was also pointed out in the work that I did with David Underdown (Talboom and Underdown 2019). However, as Sexton et al. (2018) states, the balance of trust is more than just being viewed as trusted (Sexton et al. 2017).

4.3 The approach to access

The OAIS model has been shown through the recent critiques of it outlined above to rest on certain assumptions which do not hold in practice, e.g. around the possibility of an isolated and self-contained system and of designating in advance very particular uses and users for the material being preserved. As has been stated above, the OAIS model has been very influential in shaping the perspective of the digital preservation community, but that perspective has also been shaped in other ways. In this section I consider the perspective of that community on the relationship between preservation and access and the way in which that might have been shaped by the longer tradition of thinking about this relationship which exists in the archives and recordkeeping field.

Currently the focus of digital preservation practice is more on the preservation side of things, which is understandable. But there is still a strong link between preserving digital material

and making it accessible. This can be seen in the early definition of digital preservation articulated by Hedstrom (1998), where the main objective of preservation is said to be to ensure that the material *'…remains accessible and useable'* (Hedstrom 1998). Within the DPC Handbook and the OAIS model, two important points of reference within the digital preservation community, this link is again highlighted, with emphasis put on the way that access is in large part the point of preservation (CCSDS 2012c; Digital Preservation Coalition 2015). Although there is clear recognition of this link, there is less clarity about what access means or what access to digital material might look like. There is discussion about the restrictions that can play a role in hindering access, such as copyright restrictions or other legal issues and of the tools that can support access, but there is a lack of a definition and best practices around access.

This was first highlighted by Davis (2008) who concluded that the case-by-case approach that was being used at the time was not feasible, and therefore overarching guidelines should be created (S. Davis 2008, 186–87). The AIMS workgroup came to a similar conclusion (AIMS Work Group 2012, 1–10) and again this was emphasised in *the Born-Digital Access in Archival Repositories* report, which surveyed over 200 cultural heritage institutions on their practices to born-digital access. The findings outline how there is a lack of models or case studies and no examples of policies for access to digital material, and in this case in particular for born-digital material (Appel et al. 2015).

A number of projects have been looking at access in more detail, most notably the working group that was set up by the Digital Library Federation (DLF) that looked into different levels of access for born-digital material. The produced report outlines three levels of access that could be provided and steps away from using the OAIS model as a base for producing these levels of access (Arroyo-Ramírez et al. 2020). Furthermore, the DPC themselves have extended their Novice to Know-How training to include a section on access (Digital Preservation Coalition 2021). Specifically for image and audio/visual files, the International Image Interoperability Framework (IIIF) have also been looking at ways to improve access (International Image Interoperability Framework 2022).

Access and the balance and link to be achieved between it and preservation has been a topic of discussion for many years. A concern with this topic is also apparent within the digital preservation community, albeit manifesting in new ways such as in discussion around the concept of the *Designated Community* introduced by the OAIS reference model. Currently it seems that digital preservation practitioners do want to provide access, and see this as an

important responsibility in their work, but they are unsure on how to provide it, as what is being said in the literature and reports is not lining up with what is happening in practice.

4.4 Perspectives on the role of the digital preservation practitioner

The review above has established the context of the digital preservation community, some of the assumptions on which it has been modelled and an ongoing tension in balancing preservation and access. For those working in the field negotiating the provision of access it is also about negotiating how they think of themselves as practitioners - how they think about and articulate their practice and the provision of access as part of that practice. It is through this lens on the provision of access that the perspectives of those digital preservation practitioners interviewed will now be considered.

The sections below will reflect the themes that came out of talking to individuals who are active in the digital preservation workspace. It will first focus on the difference in perspectives, as different individuals are active in this space and this results in a lot of frustration. Then two themes will be discussed where all individuals see where improvements can be made, one around the basic technical skills that are expected from digital preservation practitioners and the other around the documentation need to make digital material accessible.

4.4.1 Differences in perspectives

In Chapter 3 it was noted that although all those interviewed as Study Groups 1 and 2 could be considered to be working in digital preservation practice, differences of perspective could be perceived, particularly between those who came from a technological background and who saw the provision of access in terms of the provision of a technical infrastructure and those who were not and did not. These differences in perspective did seem to give rise to some feelings of frustration, on the part of those with a more technological focus, as highlighted by Tweed, a developer at the Wellcome Trust:

> *'It's a bunch of people that have come together to talk about a topic that they don't understand, that they have no ability to actually directly change, and so they just talk about it in circles in a really abstract way and to some degree it's really not that helpful. And you can see that changing, but that's only going to change if those groups accept the expertise that people from different backgrounds actually bring to the table.'*

This quote from Tweed emphasises the frustration felt when working with digital preservation practitioners without a background in technology and it became clear that frustration was very genuinely felt by a number of those interviewed.

All interviews started with the participants explaining their role and position within their organisation (for an overview of this information see Chapter 2). Those interviewed as Study Group 1 all described their role as either curating or preserving digital material. Pennock (British Library) and Eveleigh (Wellcome Trust) both manage a team that fulfils this function for a larger organisation. Their teams are responsible for a part, the digital preservation part, of a larger process in acquiring, preserving, and making material accessible. Moore (ADS) and Thornhill (TfL) both work in smaller organisations, where all the above-mentioned parts of the process are considered their responsibility. The same could be said for Downmunt (LCVA), but although he may be preserving and making digital material accessible, he does not think of himself in this way and talks about his practices being *'dodgy'*. Downmunt runs a community archive but currently he thinks it would be best for the archive to be *'institutionally archived'* if it is to survive for the long-term.

When talking to Study Group 1 in more detail about what their roles entailed there was, unsurprisingly, a strong emphasis on digital preservation. Pennock defined the remit of her team as follows: *'Our remit is to ensure that the Library can and does preserve its digital collections for the very long term. Which has an emphasis on preservation and not on access.'* She considered that a large part of her role was to advocate for digital preservation and the importance of it. Eveleigh explained that their remit was changing, previously: *'…the majority of the emphasis was on preservation and ensuring things were safe.'* but now they were *'…starting to see, very gradually, a demand to use it and I am particularly keen to, as we develop our born-digital program from this point onwards, to not lose sight of the purpose of preservation being access.'*

At the ADS, the remit has always included the access side, as Moore explains: *'Our ethos has always been to make data available for re-use and we see re-use as a very important part of data preservation…*'. Following on, Moore believes that this leads to a *'…much more rounded…'* preservation process as the data is then re-used by users giving an indication to how future processes around preservation may change to better accommodate users in the future.

These discussions and quotes therefore make it clear that Study Group 1 understands the link between preservation and access, and that there is a growing emphasis on access. In

their practice though they also see that they have tasks in respect of other things such as raising awareness about the importance of digital preservation internally and being able to pinpoint key players or partners to work with. The British Library is trying to achieve this by running workshops on digital preservation for all the employees and sharing responsibilities for the digital material, as Pennock explains:

> *'Our position is that digital preservation is not something that my team alone can do, it's much too big a task for that and that's the message we put out to everybody that we work with.'*

All the institutions interviewed from Study Group 1 are small groups of people within larger organisations or a small archive within a larger sector. This makes it difficult to have an impact without working with others. The ADS seek to work with others, by for example working on guides with other archaeological digital archivists, such as The Guides to Good Practice which they collaborated on with the Digital Archaeological Record (tDAR). Additionally, they work with local authorities and bodies such as the Chartered Institute for Archaeologists (CiFA) and the Digital Preservation Coalition (DPC). When talking about how to keep his archive running Dowmunt, from LCVA, points out how important it is to identify key players: *'I think that's why we are pursuing Goldsmiths and pursuing the LSA, and it's more a question of being able to identify an institution that is going to take responsibility for it.'* TfL is trying to do something similar, by working with other departments. They are currently trying to be added to Pathway, which is a project methodology at TfL, to make archiving a mandatory step in the TfL project process, but Thornhill thinks that further support is needed to make this happen:

> *'But if we can get them on board with a particular idea, it's more likely we will be listened to than if we sat there saying it by ourselves.'*

For Study Group 1 raising awareness and picking out key players seems to be considered an important element of their practice. They were all aware of the link between preservation and access, but spoke with more confidence about the way in which they preserved material rather than the ways in which they provided access to it. Even so, there was a sense that they still doubted themselves and their practice. For example Thornhill spoke of how *'I don't know how, but we have as a service somehow developed quite a good reputation for our digital approach.'* Self-doubt was also apparent amongst those interviewed as Study Group 2. Here there was a sense of having drifted or been forced by changing technology into a new role that was unfamiliar or alien. For example, as Pledge states: *'I guess more by historical*

accident than design, I'm responsible, at present, for a lot of what goes on with processing of born-digital material within our section, that's it really.' And his colleague at the British Library Foss spoke of how: *'From my point of view, just very briefly to say, I think obviously the whole digital environment has an impact on how we think about our role as archivists. We don't have any agreement on what that may look like as a profession, and that's one thing that we are grappling with at the moment.'*

Another colleague at the British Library, McKean, fully agrees with his coworkers on this: *'Jonathan [Pledge] fell into it by accident and then I fell into it by accident as well, by being relatively enthusiastic.'* Then again, Claeyssens critiqued the Dutch National Library for not having a digital curator, which led to him being assigned to the role and unsure what he should be up to: *'What has resulted in me getting that task attached to me. Currently I am trying to understand, and this sounds more positive than I actually mean to be, what this entails, this curator role. [Translated from Dutch]*'

Not everyone within Study Group 2 expressed the same feelings of self-doubt and uncertainty however. As was highlighted previously, for those with a more technological background or who were more advanced on the path of engaging with the provision of access in terms of the provision of a technological infrastructure, the more prevalent feeling was one of frustration. Returning to Tweed again, who is the Technical Product Manager at Wellcome Trust, this frustration was expressed as follows: *'But I do think, and this is obviously my own personal opinion, I do think that there is a tendency amongst some people to over-intellectualise this stuff and almost try to convince themselves that it is more complicated than it actually is. I mean I don't entirely know why that is, I could suspect that maybe it's because they are deeply intellectual people and they want to feel that what they are doing is complex.'*

What Tweed says here may sound harsh, but it showcases his frustration when trying to work with digital preservation practitioners. He is able to offer his technical skills and perhaps more importantly his perspective on what is required technically in terms of the infrastructure to provide access, but he finds it difficult to get this across to other colleagues. Nonetheless he does feel some sense of accomplishment in that he has been able to improve the infrastructure stating that: *'We just don't have those kinds of problems in those areas anymore, because we can scale to what we need, and the systems are now architected in such a way that they are faster and more robust to start with.'*

Ordelman has felt a similar sense of frustration when working on past projects expressing the sentiment that: *'But of course these scholars don't have a lot of knowledge about technology and they think: when we start a project and we get funding, yeah, that's funding for us. And it's not for…I don't want to spend all of this money on developers.'*

The difference of perspective that might lie behind this frustration has already started to be explored in Chapter 3, where a perspective was articulated that saw the provision of access very much in terms of the provision of a sustainable and interoperable technical infrastructure to enable the provision of access. In this chapter, another perspective is starting to be articulated that sees the provision of access in terms of just one of the many tasks to be performed by the digital preservation practitioner. If, as Tweed asserts his colleagues *'want to feel that what they are doing is complex'*, is the perspective in which they are right? One area of complexity in which some participants do, and some participants do not see themselves as needing *'to engage revolves around the degree to which the use to which the material is put subsequent to its being made available is or is not considered a concern to be taken into account.'* Some participants expressed a worry about losing control of the material, e.g. as Moore explains:

> *'…people always want to retain control of their program, their data, whatever it is. They are aware that once you push it out there, it becomes fixed and it is more accessible to people and people can pick holes in it and identify or use it in ways that you might not necessarily agree with.'*

What is emphasised here is the fear of losing control as individuals are able to criticise the material once it has been made accessible. Eveleigh has similar worries, especially when dealing with older collections: '…*you can download them much more easily than you used to be able to and so that's heightened our concern around the quality of the metadata that they are associated with.* This is especially true for a number of collections that may have outdated tagging that nowadays would be classified as inappropriate or racist. This highlights the fear and the responsibility that is felt towards the material that is made accessible.

Others have a very different attitude towards making material available. They do not seem to feel it is their responsibility if the material is misused in any way. As Ordelman, the developer on the CLARIAH project, states: *'…because you can always hack a system right? So you can always download stuff or you can make screencasts or videos and publish them on the internet. You cannot circumvent that.'* Tweed, developer at the Wellcome Trust, is very

much in line with this: '…*we can try to explain to you how you can re-use it, but if you break those terms and you use it differently, ultimately that is on you.'*

As Tweed continues to explain, in his opinion the emphasis should not lie on how the material may be misused, but on how the material is made available and under what licences. This information is important to get across, but Tweed does emphasis that that is everything that an institution is able to do.

This difference in perspective, in where the line is drawn around the responsibility for the use to which material is put, manifests in a difference in the level of caution or risk appetite practitioners expressed in relation to the provision of access.

Those who were less cautious also expressed the view that everything did not have to be perfect for access to be provided. Ordelman, Sherrat, Corti and Izzo all mention this in their interviews. For example, Sherrat spoke of how: '*I would rather have institutions just get stuff out, rather than think they have to clean it up, or fix it, or whatever.'* Corti agrees stating that: *'You know, I agree, having something quick and dirty is really important because the whole point is they are not there to use that thing, even if it is a PDF or an image, at least you can see what it looks like. I think quite often descriptions can be rather frustrating and we could do a bit more in exposing things up-front, so that people can see what's there.*

One of the findings to emerge from the interviews with Study Groups 1 and 2 was of a difference of perspective within these groups. This difference and the different perspectives it engendered (between a view of the provision of access in terms of the provision of a sustainable and interoperable technical infrastructure and a view of it in terms of this being just one of the tasks to be accomplished through digital preservation practice) are now being explored and articulated in more detail in this chapter. Despite the fact of this difference in perspective and the finding that it can on occasion cause feelings of both self-doubt and/or frustration as outlined above. Additionally, the interviews gave insight into how it could be manifested more positively as an opportunity for those involved in the practice to both develop new skills and strengths themselves whilst acquiring a greater appreciation of the complementary skills and strengths of others.

4.4.2 Basic technical skills

Nonetheless there does seem to be a feeling, as expressed here by Tweed that there is now a need for all those working in digital preservation practice to have some basic technical skills or understanding, not having them is: *'…just not okay anymore when they are dealing with born-digital stuff.'* During the interview at the British Library, McKean mentioned that he has

been able to follow a computing for cultural heritage course at Birkbeck University giving him some experience with scripting languages, which he has been able to use in his work. He may not be an expert, but he is able to understand the basics of it. Sherratt, from the GLAM Workbench, spoke of how he taught himself how to use Python and that one of his main goals in creating the GLAM Workbench is to encourage others to do the same: '*I mean primarily it's aimed at researchers, to recognize what they can actually do with GLAM collections, to recognize that there are different ways to do research now because all of this is now becoming available from GLAM organisations. Getting them over those initial barriers so that they can start seeing what's possible and then carry on from there.'* Milligan, from Archives Unleashed, is very similar to Sherratt in the sense that he was able to teach himself, using courses such as The Programming Historian to become able to make Archives Unleashed successful.

If the consensus seems to be then, that it is both helpful and necessary to have some basic technical skills to work in digital preservation, this does not mean that non-technical skills are no longer recognised as important. Ordelman, from the CLARIAH Project, finds collaboration really important and mentions this when working together with historians and researchers, who are essentially the users of the tools that have been developed: *'They may not be technology people, but they have, of course, brain power that we can use to really think hard how we should, in co-development, develop this, quite complex infrastructure.'* The Technical Lead at Project Alpha mentions the same; that they see the value of working with people from different disciplines.

The discussion of the development of technical skills sometimes seems to be connected or associated with the topic of automation and with a concern that technology will render human intervention obsolete. This concern is recognised but not thought valid by those more engaged on the technical side. For example, Tweed speaks of how he views technology as a tool or an amplifier and sees the automated techniques not as a replacement of archivists, but as a way to make them more effective: '*An archivist doesn't need to create a list of files, an archivist doesn't need to describe what those files are, an archivist does not to try to document what's inside them. They need to tell me why that particular set of files is important, not what's in them.'*

For those less engaged on the technical side though. This concern around automation seems to relate to that, previously discussed of losing control over the material. Eveleigh worries about implementing automation too quickly: *'I think leaping from completely manual*

processes to completely automated is really dangerous because you have no way of troubleshooting where it goes wrong, so we're trying to take that slowly and gradually add in more and more steps to be able to make that process.'

Moore, when talking about their automated system called ADS-Easy, talks about similar concerns:

> *'Good in a sense that it improved your workflow, there is less human interaction with data, less human involvement in it. But that's sometimes a good thing and a bad thing really, because invariable the computer is only an automated system and is only as good as the instructions that you give it.'*

Then again it seems to be associated quite strongly with a concern with ideas of description and documentation being incomplete or incorrect when done by an automatic system. This concern is raised by Pledge, who talks about cataloguing PowerPoint files with no way of discovering what the content was as a manual and laborious task: *So I went through and copied out the first line of all of the PowerPoints, so you had the title of the PowerPoint, or that presentation at least, to give you some clue. And that was obviously a very laborious task to do. Most of the metadata we require is technical metadata at present, and at scale that's the best we can do unless we come up with some sort of scripting solution to acquire further descriptive metadata.'*

This is in contrast to what Tweed talked about above when mentioning automation. Tweed does not see this task as crucial for the digital preservation practitioner and finds it more important that time is spent doing other tasks of importance. What this mainly showcases as well is that there is a lack of understanding in what technical skills may be needed, but communication seems to be key here.

4.4.3 Documentation

Thinking about documentation and description does seem to emerge from the interviews as a non-technical skill that is very valuable within the digital preservation space, but a need for change is also recognised here too. One of the main benefits of making material available for computational access is that it is available in bulk, either by requesting certain data from an API, outputting data from a platform or downloading pre-processed datasets. This obviously has an impact on the form of documentation that should accompany the material. In the interviews with participants, some insights into what should be provided with data became clearer as a number of participants mentioned that documentation is currently inadequate

from both institutional perspectives (as highlighted by Tweed and Izzo when talking about their APIs) and user ones (as highlighted by Milligan and Sherratt who would have liked to have more information provided with the available archival data that they have found online).

Documentation is now needed not just to identify the digital material itself but also to explain how it was processed. An example of this is given by Sherratt, of when he used a dataset and was unclear about it - only his wife was able to explain the disrepancy to him as she was an expert on the subject matter. Sherratt asks for clearer documentation on processes: *'Obviously if those processes where data is transformed in different ways can be documented, it makes the data more useful and it also avoids people making mistakes, because they just don't know the history of it.'*

Cleayssens from the Dutch National Library sees this need to provide such documentation around processes as one of his priorities as a digital curator and he is currently writing a history on the digitised collections of the Library. He sees this as especially important for digital material, as a substantial amount of choices are made during its processing: *'….And I hope to, within a couple of years, publish the history on our corpora because I am noticing that researchers are starting to use our material more for their research. This material has a number of hidden selection processes, hidden technical elements, you could even say hidden social elements. Currently we give little information about this, while this is even more important to provide for digital collections.'*

Then again Dethmers hopes to offer similar documentation for the HathiTrust, as he currently sees this problem as well, he is concerned about biases and gaps in datasets not being signposted correctly. He talks about this in the context of individuals looking at doing data-driven research: '*Okay, all these materials came from that particular research library, there is unlikely to be a certain subset of materials. Like we don't have grey literature, we don't have archival material, really. So it's important to know that when you are doing data-driven research, because there could be gaps in your dataset and biases.'*

The nature of digital material does therefore seem to call for '*…more complex documentation, particularly perhaps in respect of the need to document'* as Milligan proposes. The Digital Scholarship Librarian at HathiTrust talks about how they have implemented versioning after running into a couple of problems with their available datasets. An example given is when material gets rescanned, and a hand or other mistake may disappear from the original digitisation image:

> *'I don't think this is an issue for the general researcher, because if there is a hand on a page and it's gone, you're happy right? If you are a scholar doing text or data mining and you download data on June 1 and some pages are not great, you know the OCR is not great, and then someone comes along in November and wants to recreate your work and they sync the same volumes, the same text, they are going to end up with different data. And that's a problem that I don't think we have fully grappled with and it's one that for more researchers could be an issue.'*

The provision of adequate documentation is therefore highlighted as another perspective on the provision of access and its provision would seem to be seen as one of the non-technical skills required to undertake digital preservation practice. Nonetheless there is a strong sense that all those involved in digital preservation practice should have some basic technical skills and understanding.

4.5 Working in digital preservation

This chapter has explored how the provision of access is being negotiated through the lens of the roles and responsibilities of digital preservation practitioners. In so doing it has explored in more depth a difference of perspective first highlighted in Chapter 3. This difference has now been further articulated as a difference between seeing the provision of access in terms of the provision of a sustainable and interoperable technical infrastructure and seeing it in terms of it being one of the many tasks to be completed by the digital preservation practitioner (alongside others such as advocating for the need to take the preservation of digital preservation seriously).

To support this articulation, the chapter started by establishing the context of the emergence of digital preservation practice as a distinct community, brought together via the common model of OAIS, but increasingly questioning some of the assumptions of this model and hence also the assumptions on which their practice has previously been built. Assumptions which were being found not to hold in practice were identified in particular around the possibility of an isolated and self-contained system and of designating in advance very particular uses and users for the material being preserved. The chapter has also considered thinking from the archival tradition in terms of a shifting the emphasis between preservation and access as another context of relevance.

For those participants working in digital preservation practice the negotiation of this tension could be seen to give rise to a variety of feelings from self-doubt and uncertainty to

frustration. This frustration was felt particularly by those more aware of the perspective of the provision of access as the provision of a technical infrastructure. From this perspective in particular there was a desire that all those practising digital preservation should develop skills and knowledge more directed towards this perspective, but there was also an acknowledgement that more was needed than just this. To be sure the provision of access could be seen as just another task for the digital preservation practitioner to accomplish, but its accomplishment was not just a matter of the provision of an appropriate technical infrastructure, but also the provision of adequate documentation. Practitioners could have strengths in making provision of one kind or another, but they also needed to have a basic understanding of what was required for both.

In this way the theme of working together that was also noted in Chapter 3 has been elaborated, as well as the idea that the provision of access is not just a question of technology. These themes and ideas have only been elaborated however through the lens of the digital preservation community and its practice, not at any other level beyond that and it is to this that we turn in the next chapter.

5 Society and the digital

The previous two chapters have explored how the provision of access is being negotiated by digital preservation practice with the articulation and elaboration of three interconnected perspectives: first, the provision of access as just one of the tasks to be accomplished by digital preservation practitioners; second, the provision of an appropriate technical infrastructure; third, the provision of appropriate documentation. The overarching frame for all these perspectives has been set however by digital preservation practice. In this chapter the intention is to open out that frame, by starting from the perspective of the broader digital environment/online public space. To assist with this, perspectives will be introduced from the interviews undertaken with Study Group 3 - data journalists.

5.1 Big Tech and the online public space

The online public space where digital material is being made accessible has been recognised as being problematic in a number of ways. For example, Yeo (2013), who referred to this online public space as *'cyberspace'* raised the issue of trust and discussed how it is easy to disseminate misinformation in this space. He highlighted how difficult it has become for digital repositories, and especially memory institutions, to embody trust, as trust in professionals in general is waning in this online space and there is also a problem of disintermediation (Yeo 2013, 216–17). This point has been raised by The Royal Society which highlight that the way in which information is now spread in the online space enables mistrust and the wider spread of misinformation (The Royal Society 2022). Questions continue to be asked about the nature of the role for digital repositories have in this space: *'…what role is there for a library when these tools are being created by trillion-dollar industries?'* (Lankes 2019).

Since the late 2000s, web infrastructure seems to be dominated by Big Tech. Big Tech normally refers to the Big Four, the four largest tech companies, Alphabet (Google), Amazon, Apple and Meta (Facebook) (Mayer-Schonberger and Ramge 2018; Alcantara et al. 2021). Other definitions talk about the Big Five, which includes Microsoft (The Economist 2018) and recently there is discussion of other large tech companies being classified as Big Tech, for example Twitter and Netflix (The Economist 2018; Levine 2021). Whatever their number, it is undeniable that these tech giants have a significant influence on how the online public space is configured, as will be discussed below. In this space, users engage primarily via the platforms generated by these companies, these platforms are *'…digital infrastructures that enable two or more groups to interact'* (Srnicek 2017, 43).

The growth of social media platforms has made initiatives such as Archives 2.0 possible, but it has also increased fake news. Fake news, or disinformation, is not new and has been around for centuries, however the abundance of it is new and social media platforms have worked as enablers by encouraging users to spread fake news as a money-making strategy (Guess, Nyhan, and Reifler 2020).

As outlined In Chapter 3, digital preservation practitioners have been and continue to be engaged in taking advantage of the new tools brought about the evolving Web and its platforms. Many of these platforms are run by companies whose main motivation is a profit one, but it is these platforms that are influencing users and shaping their expectations of the online environment. Lanier (2018) has classified some of these platforms as Bummer*: Behaviours of Users Modified, and Made into an Empire for Rent* and speaks of how such platforms can be harmful for individuals (Lanier 2018; 2019). Of course, good things can still be performed on Bummer platforms, but their problematic is ever-present.

Another example of how the platforms being built by Big Tech are influencing the expectations of what access looks like in the online public space is the development of keyword search. A survey conducted in the Netherlands on how academics search for digital material, summed it all up with the phrase *'Just Google It'*, as this was the most common digital search method (Kemman, Kleppe, and Scagliola 2012). Furthermore, a large number of users come into contact with memory institutions for the first time through Google, for example, the Europeana portal indexed their material and saw a huge increase of pages visited (Nicholas and Clark 2015, 28). However, users finding their way into catalogues and portals created by repositories can be confusing and make it difficult to understand what they are looking at.

Related to users coming in through search engines such as Google is also the influence that the keyword search box that Google has on the development of tools within memory institutions. Currently there is a lot of debate around the usefulness of this approach. As Gilliland (2016) and Winters & Prescott point out, this simple keyword search may not be enough to actively engage with the digitally preserved material (Gilliland 2016; Winters and Prescott 2019). Furthermore, Putnam (2016) and Romein et al. (2020) warn about the pitfalls of using search engines, as they are not transparent and can be biased, influencing the research that is being conducted (Putnam 2016; Romein et al. 2020). Guldi (2018) does come with a part solution for this by introducing a term called 'critical search' which encourages researchers to use critical thinking to question the algorithmic output of, for example, these

search engines (Guldi 2018). However, this is only possible if the search techniques are disclosed by the institutions, which is currently not the case.

Another more indirect way in which Big Tech platforms are influencing the experience of access in the online environment is by the facilitation of context collapse, where multiple audiences are flattened into a single context. An example of this is a wedding, introduced by Joshue Meyrowith in 1985, where the bridal couple brings coworkers, friends and family together and it becomes unclear for them how to communicate and behave in that setting, as they may act differently in a professional setting than with friends or family. This theory uses electronic media, especially broadcasting, as the example of this phenomenon:

> *'The combination of many different audiences is a rare occurrence in face-to-face interaction, and even when it occurs (at a wedding, for example) people can usually expect the speedy resumption of private isolated interactions. Electronic media, however, have rearranged many social forums so that most people now find themselves in contact with other in new ways. And unlike the merged situations in face-to-face interaction, the combined situations of electronic media are relatively lasting and inescapable, and they therefore have a much greater effect on social behavior' (Meyrowith 1985, 5).*

Marwick and Boyd expanded this idea of context collapse into a social media setting (Marwick and boyd 2010). They talk about how the imagined audience online may be very different from what the actual audience is (Marwick and boyd 2010, 123; J. L. Davis and Jurgenson 2014).

As well as context collapse, another issue in the online public space is a temporal collapse, which is visible in many social media feeds (Brandtzaeg and Marika 2018). Burgess (2019) is especially worried about this type of collapse for libraries and other similar institutions (Burgess 2019), this concern is also echoed by Odell, who talked about how her library experience of trustworthy material is the complete opposite of the online public space, and specifically the news feed that many of these platforms offer: *'Nothing could be more different from the news feed, where these aspects of information – provenance, trustworthiness, or what the hell it's even about – are neither internally coherent nor subject to my judgment. Instead this information throws itself to me in no particular order, auto-playing videos and grabbing me with headlines. And behind the scenes, it's me who's being researched'* (Odell 2019).

These problems around context collapse and the influence that the platforms, and especially Bummer platforms, are having on memory institutions have made this online public space a hostile environment for digital preservation practitioners to operate within. Especially when considering that the main goal of these platforms does not align with the memory institutions who promote trust and provenance.

Furthermore. a recent report by Lankes (2019) described how Google is more trusted than libraries and most elected governments and yet it could also be argued that companies such as Google are undermining truth by embedding unhelpful (except in a profit-making sense) patterns of behaviour and ways of working into the infrastructure of the Web. Lankes (2019) has articulated that: *'However, the biases we bring, or more precisely the principles, we bring to the Google and Facebooks of the world is that a strong voice that advocates for transparency, privacy, the common good, and a need for a durable memory is important.'* (Lankes 2019). The question of whether and how memory institutions can be this voice remains an open question.

One issue on which digital preservation practitioners have started to try to raise their voice, with some success, in recent times is around the lack of transparency. Many seemingly simple search boxes are underpinned by complex algorithms and Guldi mentioned a method called 'critical search', which advocates for researchers to use critical thinking to question the algorithms used in these searches (Guldi 2018). Then again from a more technical side, Andresen has explored if it is possible to translate the algorithmic output into a more meaningful format when making material accessible (Andresen 2019). Additionally, there has been an exploration of explainable artificial intelligence within the digital preservation context, which explored the possibility of explaining the black box algorithms increasingly being employed, often invisibly (Ridley 2019; Bunn 2020).

Furthermore, digital preservation practitioners are starting to engage with the issue of economic dignity. As Lanier explained, certain tasks, such as translating of text, have become jobs that are seen as obsolescent by Big Tech. This is because they have pushed for the assumption that these translating tasks can be performed by AI systems. But these algorithms need training data, which is generated for free by people on many of the Big Tech platforms (Lanier 2019, 135). Questions can be asked about how best to surface and reward this labour and past experiences with crowdsourcing and Web 2.0 means digital preservation practice can speak with some knowledge on such questions. Web 2.0 was seen as a way to democratise content online, as anyone could create content, and it was not limited to big

institutions and news outlets (Srnicek 2017, 53), but again, even before Web 2.0 evolved, Terranova highlighted the assumption inherent within in that users generate free content (Terranova 2000).

This review of some of the issues arising in the online public space point to it being a relatively hostile environment for memory institutions to operate within, e.g. in respect of the provision of access. Big Tech, one of the key players in building the infrastructure of the Web, have embedded certain expectations and patterns of behaviour into core values and practices and these do not align very well with the way in which memory institutions wish to provide access to the digital material they hold.

5.2 The Data Journalists

Another group operating alongside digital preservation practitioners and memory institutions in the online public space are data journalists. Data journalists share similar core values around trust and provenance with those who practice digital preservation. They may not preserve digital material, but they do provide access to it in the online space. This section seeks to introduce and contextualise data journalists by providing a brief overview of the development of their community and practice It then elaborates on that perspective by reporting aspects of the interviews carried out with them (as Study Group 3). This will be highlighted within the following themes: hostile environment, pushing back and influence of Big Tech.

Data journalism, if seen as the practice of improving journalism with data, dates back to the early 1800s. Then again, if considered as the practice of improving journalism with the help of digital data, it can be dated to a famous 1952 example of the prediction of the results of the US elections (Gray, Chambers, and Bounegru 2012). From these early origins, data journalism started to professionalise in the 1970s, when open social science data became available, but it was not until the 1990s when it really started to take shape. The first handbook on the practice, then called computer-assisted reporting, was published in 1996 and is now in its fourth edition (Houston 2015a). The blossoming of the practice was partly due to a growing number of seminars run by The National Institute for Computer-Assisted Reporting (NICAR) and the Investigative Reporters and Editors (IRE), making it possible for data journalism to be shaped into its own discipline (Houston 2015b).

In 2006 Holovaty wrote an important blog post critiquing the current approach of the journalism sector towards technology. He argued that this approach should be about giving readers access to the data used within articles, not necessarily about implementing the

newest technology, such as making websites functional on a mobile phone (Holovaty 2006). For many this is seen as the first time that data journalism is articulated in the way that it is used today, as explained in The Data Journalism handbook, one of the key texts within the discipline (Gray, Chambers, and Bounegru 2012). A few years later the first conference on data journalism was organised in Amsterdam (Kayser-Bril 2010).

This was the point when data journalists started to see digital material as an entity in its own right, not only as a tool in their story, but also as something to which access should be provided as a part of that story. Furthermore, wider external factors played into this shift in attitude, with the growth of openly available data, for example through data portals and new technologies making it possible to do more with this newly available data (Gambini 2019).

Data journalism manifests in the rise of fact checkers and with them an increased interest, on the behalf of data journalists, with educating the public and other journalists in data provenance and trustworthy sources of data (Guess, Nyhan, and Reifler 2020). Heather Krause, who founded Datassist, has many useful tools and guides on her website focusing on provenance and documentation of material collected by data journalists (Krause 2019b; 2019c; 2019a). Another example of this is the work of Winny de Jong on the Data Journalism Tools which offers an overview of tools to help analyse and visualise data (de Jong 2021).

Currently data journalism is slowly becoming part of the profession of journalism as a whole and is not seen as a sub-part anymore. There is a strong emphasis on developing data skills within the profession, but also with educating the general public on data literacy (Barr, Chalabi, and Evershed 2019). On this issue of the importance of data provenance and literacy, data journalists share common ground with digital preservation practitioners. There is common ground in the way that both groups are concerned with getting it right when anything is published or made available. As Vanetta highlighted in an article written by Houston: *'In news, you can't make mistakes – there is a reputation to take care of. The editorial team is not as used to failure as developers are'* (Houston 2015b). The data journalists interviewed had a lot to say about their role and responsibilities and their perspectives will be elaborated below.

One of those interviewed, O'Donnell, teaches data journalism at Cardiff University, whilst another, Tarling works at the BBC and two of those interviewed worked for newer organisations which would be classified as fact checkers; Bellingcat and Full Fact. For the fact checkers in particular, Bellingcat and Full Fact, the provision of access to the data used to write a story was considered particularly important because it was felt that in order

to uphold their reputation they had to use an approach where everything was open and transparent, and anyone could follow their work back to the source. In the interviews with data journalists, many themes and trajectories were discussed that mirrored those raised by those working in digital preservation. Those, like Tarling who had decades of experience in their practice, were able to reflect on and articulate a distinct digital shift, e.g:

> *'The more general problem of accessing information, I think what I have seen change is the move from physical access and paper-based systems, and meetings, and exchanges of documents towards much more machine-based access. And now we have people who are actually employed as data journalists, those sorts of journalists are journalists who have a degree of technical skills.'*

As was reported in the context of digital preservation practice, some journalists had become more specialised in working with digital material. There was a similar feeling that, some basic understanding of technical skills is a requirement, as Tarling puts it: *'This is not to say that they are software engineers or developers, but I would expect them to be fairly capable with unstructured data and to be comfortable with tools like Microsoft Excel for example.'*

Furthermore, Tarling sees the future data journalists not as a developer, but as a data journalist with basic technical skills that is able to work with more technical individuals such as data engineers: '*Now there is a sort of line that you come to which is where you are moving outside of those sort of, almost like kind of advanced office-based skills to data manipulation, and what we've found is that for more modern data journalism it actually pays to pair up a data engineer or data scientist with a data journalist.'*

This is very similar to the ideas proposed by the more technical individuals in Chapter 4, where digital preservation practitioners are able to use some basic technical skills, but if they need something more advanced, they are able to communicate with more technically minded individuals. Tarling summarises this as: *'…it's a mixture of some basic technical skills and a kind of mindset to enjoy querying data and looking for interesting inferences.'* He also highlighted how it is important to be able to communicate with the more technical people within the institution.

No one involved in digital preservation education was interviewed as part of Study Groups 1 and 2, but this perspective was present in respect of digital journalism in Study Group 3. O'Donnell, who teaches data journalism to bachelor students in journalism, stressed how

important it was for all journalists to get an introduction to data skills, even if they didn't end up specialising in data journalism. He spoke of: *'…an obligatory, which is probably the key word, second-year module on data journalism, theory and practice. So I teach that with a colleague and then we also have the MA programmes.'*

O'Donnell finds this of importance as he is of the opinion that not everyone on the degree needs to become a technical individual but needs to be aware of the digital as a resource. He also prepares them to be able to work with the minimum as he is trying to prevent future journalists from leaving the room when someone mentions something that is either numerical or computer related. It seems to be about building confidence and communication in digital skills.

The common ground between data journalists and digital preservation practitioners in respect of feeling a responsibility to get things right in the provision of access has already been highlighted. O'Donnell elaborated on this in terms of data journalism being a perfection-based discipline as follows:

> *'The point has been made that journalism is basically a perfection-based discipline, whereas software engineering and technology is iterative. So technologists, what they do is they build something and if something doesn't quite work, or something that works and they build a better version. That's why you have "Word 26.2". Whereas if journalists get it wrong, if they publish something on the front page that is wrong the first time, then they go to court.'*

There were though perhaps slight differences in that the emphasis for data journalists seemed to be on reputation, their own reputation as an authoritative source, whereas for digital preservation practitioners the emphasis seemed to be more about facilitating the responsible use of digital material and ensuring it is set in its proper context.

Another area of common ground was a shared concern with good documentation. For Higgins, who works at Bellingcat documentation is seen as one of the most important tasks and falls under the core of what they do: *'…identify, verify and amplify…'*. For Higgins verifying and documentation is seen as important because Bellingcat are open-source investigators and have to be able to prove how they have come to their conclusions. Furthermore, the process of documenting material is seen as important by Benedictus, who again considers it important to explain what process was used to come to conclusions.

O'Donnell sees this as an important step as well and teaches his students to save a copy of any source that they are using, especially as data sources can change in quite quick timeframes.

Data journalists have realised how important versioning can be for their work, which is also highlighted by a number of digital preservation practitioners in Chapter 4. O'Donnell talks about how important it is to save the source, as datasets can quickly change and it is important to have a copy:

> *'You know, some of the stuff is updated literally like every hour. Anyway, so what we do, we make sure that they have a copy of what they did. It sounds really grand, the way that I have explained it to you. But it is also obviously a teaching device, because it means that you can be a lot more free and easy, and make mistakes and delete things. And you don't have to be precious with what you are dealing with. You get questions like: Should I hide columns or delete rows? Or whatever. And you are like: doesn't matter, do what you want, because we have a copy.'*

Finally, another skill that is seen as useful by O'Donnell, is what he likes to call interrogative visualisation. This is basically a way of representing material and then questioning that material, as O'Donnell explains: '*But data journalists, all the time, visualise results. But there is also this aspect of interrogative, well I call it interrogative visualisation. It is what scientists do all the time. Like graph my results to see if there is anything in there.'*

This section has introduced and reported the perspective of another group who operate within the online public space and who have some interest in the provision of digital material (in this case the material on which and from which their journalism is based) within that space. This group was seen to share much common ground with digital preservation practitioners. Both groups seemed to consider that the digital shift had led to need to develop a basic understanding of technology and data and to place great emphasis on data provenance and literacy. In this context both groups shared a concern with good documentation and versioning and with providing access to material in ways that would promote trust and accurate understanding. Both groups therefore seem to share a concern not just with the provision of access to digital material, but also with the reception and subsequent use and interpretation of that material by others.

5.3 Perspectives around the online public space

Elaborating on these areas of commonality and difference between the data journalism and digital preservation perspectives, this section will consider such commonalities and differences in more detail, starting with the commonality that, from both perspectives the online public space can be seen as a hostile environment, hostile that is to their core values and to what their practice is about.

5.3.1 Hostile Environment

Within the interviews with data journalists there was concern expressed around social media platforms. Bellingcat relies on data that comes from social media platforms and Higgins spoke about how this material is used as evidence not only within the articles that Bellingcat publishes, but also in court. Higgins worried about how most of the public did not understand that they are effectively publishing material on these platforms and that this publication could have consequences.

Tarling spoke of how this ease of publication is making it more and more difficult to know what comes from reliable sources and what does not and he highlighted how advertising (and hence the profit motive of many platforms) exacerbated this issue, e.g.:

> *'In that environment that we live in now, it is very difficult for publishers to convince consumers that this is legit, so the rise of fake news as a sort of inevitable response to the presence of the platforms between the publisher and the consumer, I think.'*

Tarling goes on to explain that it is especially difficult as these companies have a *'…vested interest in the existence of fake news.'* This vested interest is because of the monetisable nature of the fake news. Similar to the digital preservation practitioners, it is difficult for the journalism field to portray the brand values of their organisations, Tarling mentioning '…trust and provenance…' as two of these values. Especially when considering their material is changed and shown alongside adverts. Tarling worries about how these values may be portrayed within these platforms: '*How do we convey those values and that truth and veracity that we have at source, how do we get it through to an audience in the age of the platforms?*'

As was explained within the first section of this chapter, this online public space of the internet is a hostile environment for institutions that have different core values around trust and provenance and it can be difficult to portray this to audiences on these platforms.

5.3.2 Pushing back

What seemed to be most important to the data journalists were questions of transparency, truth and trust. Higgins, from Bellingcat, spoke of how: *'…one of the paradoxes of open-source investigation is that you don't have to have a reputation to show something is true. You can demonstrate step-by-step.'* Higgins sees this as vitally important as: *'…the techniques that we have developed means that we can break down step-by-step how we have come to our conclusions. So that is something that we have always really done with our work, if we are saying something we can explain why we are saying it and often we do that in the article, we explain step-by-step how we came to that conclusion.'*

Full Fact is completely in line with Bellingcat, as Benedictus explains how he does not assume that readers trust the material that is being shared with them. Instead of stating certain facts, such as being politically neutral, they showcase this neutrality by maintaining a certain process in their work. Benedictus goes into detail about this process:

> *'And we will use those data sources to justify our conclusions and then we will link to those data sources. So, in theory, anyone who looks to any of our articles can reconstruct the whole process of writing it themselves by using the links that we provide, and that is how they can see us as trustworthy, provided they trust the original source of the data as well.'*

The BBC is different from these two organisations, in the sense that their processes and sources are not as transparent. Rather they put a lot into the hiring process; working on the basis that these checks ensure that anyone who has been hired can be trusted to do their work correctly and thereby taken on trust.

O'Donnell critiques this approach and for him transparency is demonstrated more along the lines of the approach taken by Full Fact and Bellingcat. Indeed, he uses Bellingcat as an example in his explanation of this:

> *'Look, it doesn't matter what you think of us, here is our method. Here are our sources, this is what we did. And it's a model that's very close to what universities would do. It's your standard transparency, you cite a piece of research, you give the details so people can look it up.'*

Showcasing this transparency in practice can be challenging, but O'Donnell has found interactive notebooks an aid in this process of transparency and uses it during his teaching and praises the ability to portray data alongside text and sees this as a crucial step to

transparency: '…*you need to have your workings, and your data and your findings super accessible and not just accessible, but also super understandable.'*

Coming back to the transparent process, Benedictus finds this to be in line with his idea of not presuming that you will be trusted, but always striving to act in a trustworthy manner:

> *'It's not sort of saying: we are The National Archives and we are great, and we are there and we are thorough and we don't make any mistakes, you can just trust us.*
>
> *No matter how true that might be, it's not something that you can prove, and you might be wrong. So the best you can do is demonstrate trustworthiness, which is: We will be incredibly transparent and show you everything that we can possible show you, we will admit it when we get things wrong and publish corrections, we won't try and hide the fact that we've done that and we will give you all the evidence and sources that we possibly can to show you where our things have come from. So actually the best way to be trusted is to be trustworthy in the way that you behave and that is what we do.'*

This again emphasises that point around being a trusted organisation not being seen as enough, it is also about acting in a trustworthy manner. Furthermore, another technique that Bellingcat uses to ensure transparency and trustworthiness is the use of storage by an external entity, to ensure that they can guarantee that they have not tampered with any of the material. This is especially of importance for the evidence that Bellingcat holds. Higgins talks about how by giving away that control they are considered to be more trusted: '…*by giving over the control of the archiving, we are doing more to verify if it is genuine and that is probably what we would use more. Because if we are archiving stuff and it's something that someone is going to deny, but it's not something that is likely to be used as evidence, then there is no reason for us to do this kind of evidence preservation, offline using our own systems, because we will just be accused of changing it anyway by the people who are denying it.'*

The mention of archiving is of interest here in highlighting the difficulty of ensuring sources can ultimately be trusted. There is a sense then that although the data journalists may not see this as their problem, their responsibility being more to document the process/steps taken from source to conclusion, there is nonetheless a shared concern with data

preservation practitioners with speaking up for trust, truth and transparency in an environment that seems hostile to all of these things

As well as wanting to act transparently and in a trustworthy manner, data journalists are also concerned with acting both legally and ethically. One example of an area where acting ethically can be complicated is web scraping, which O'Donnell explains to be a bit of a *'grey area'.* He informs his students to at least check the terms and conditions and that most of it comes down to ensuring you are not banned from the servers.

Furthermore, there was discussion from those data journalists interviewed around legal considerations. For example, Higgins worried about laws that have been put into place by governments who do not understand the full extent of open-source material. Higgins gave the example of hosting terrorist material on archival websites; by law these archives could then be seen as terrorist sites, but they are distributing this material with other intentions in mind. Higgins offered a solution to this issue in the form of an index, where people can showcase what material they have, but not actually share it.

The online public space can perhaps be seen as a hostile environment then, not just in terms of the patterns of behaviour and expectations set and embedded into that space by the Big Tech players who constitute a large part of its infrastructure, but also in the sense that legislation has not kept up or developed in pace with these developments - making it hard to engage in the new space and continue to act legally.

This issue was definitely felt by those interviewed from a digital preservation perspective, where the legal framework around copyright was seen as a particular problem. For example, the issue was highlighted by both The British Library, where material is only available in the reading room, and The Dutch National Library. Then again Ordelman at the CLARIAH project spoke of how the existing legal framework is making it difficult when making computer-generated material accessible. He ran into several problems when trying to make it possible for users of the CLARIAH platform to download their outputted data and he thinks this mainly comes from misconceptions that are currently in place around AI and the use of algorithms:

> *'But for us, we are the opposite, we just want to provide good ways for researchers to do their research. So when we want to provide, for example, export tools to export word frequency lists that are generated using some kind of NLP tool upon the metadata, then there will be, there*

will probably be legislation that is not allowing us because of the problems that exist in the commercial world, let's say.'

Ordelman finds that current copyright laws are outdated; he views computer-generated data as new data, whereas the current copyright laws see this as a derived dataset from in-copyright material, constraining the use of the created computational dataset. The UK Data Service, who have been making data available for decades now, have thought about the legislation issues around this material in detail, as they had to, but they certainly don't underestimate the work that goes into that. As Corti explains, legislation is used as an enabler, not a constraint: *'And it has to have a legal gateway, so we can use the GDPR as a legal gateway or the Education Act or the Digital Economy Act, or the Statistics Act, it has to have a gateway.'* HathiTrust have come up with a non-consumptive approach, which made it possible to use in-copyright material for research, and the Dutch National Library may be exploring a similar approach if they keep having issues around copyrighted material.

Those working in digital preservation are finding ways of working under existing legislation, but it is not proving easy. For example, both Wellcome and the Dutch National Library talk about how they initially underestimated the copyright issues. Claeyssens said: *'…that aspect is maybe the thing that was most underestimated by the Dutch National Library when we started doing mass digitisation and in general for our digital collections.'* They now have a number of specialists in place and ensure that they make future agreements not only to make files accessible on an individual level, but also in bulk, for computational access, as this has caused numerous problems in the past.

5.3.3 Influence of Big Tech

Another commonality that was introduced in the previous section was that both data journalists and digital preservation practitioners seemed to share a concern with the reception and subsequent use and interpretation by others of the material to which they are providing access. These then are further terms in which the provision of access can be viewed. This concern with reception brings others firmly into view and these others are, for those in digital preservation practice at least, often spoken about in terms of users.

When talking to Study Group 1, many interviewees expressed uncertainty around how users should be approached and accommodated, especially in the online public space. There was discussion around understanding the user and knowing who is accessing the material being made available. As Eveleigh pointed out, in the analogue space archivists actually met their users and could gauge in person their reception of and reaction to the material being made

available to them, since those users were only those *'…brave enough to come through the doors.'*

Moore from the ADS highlighted that they are finding it increasingly difficult to understand their users in the online public space. There are a number of uses pointed out by Moore of things that they would have never expected the material to be used for. In his view this makes it even more challenging to recognise what community is using their data. They serve an archaeological community; this community works in both the commercial sector as in a research environment. However, since making their material accessible, different groups have been using their material for different purposes, as Moore elaborates:

> *'…we also get a huge variety of users who work beyond that scope or beyond that group. People who are interested in their local area, people who are interested in a facet of the past or heritage that they might be interested in, and increasingly people interested in crafting and things like that and there are resources available there that people can use data for that. And even really strange re-use, depositors who have produced 3-D models and people have printed them out and made kind of paperweights from them. Stuff that you wouldn't really expect.'*

Then again concern was also expressed around the tools, in the form of the online catalogues that archivists have published on the Web and how they are perhaps not in a form that works very well through that medium. As Eveleigh pointed out: *'I think often the problem is that people, people's first point of contact is via Google, so they land on a random page and then find it very difficult to navigate within any context.'* Nor are they necessarily familiar to navigate around and understand these pages as Moore from the ADS explains when mentioning the faceted search for ADS material:

> *'For some users, it is quite challenging, you need to understand how to use it and what you're searching for to really get the most out of that and that requires time and effort. People, invariably, are so used to having an environment where they just send in a term and they get a series of results back, that is not what they expect.'*

Previously it has been identified that one perspective on the provision of access taken by digital preservation practitioners negotiating this access, is to see it as the provision of adequate documentation. On these terms it would seem that some have doubts as to the

adequacy of what has been provided so far. Eveleigh, for example, wonders if catalogues are actually the best approach as: '*…professional descriptive standards in any field have not been developed with the user in mind.'* Dowmunt, who runs a community archive, is perhaps more hopeful and talks of how: *'…the idea of the site is that you can access it at any point and be pointed towards other relevant material if you wanted to and there are multiple ways of searching it. You can search by year, by theme, there are different ways of doing that.'*

Consideration of users also perhaps start to highlight one last perspective on the provision of access that digital preservation practitioners are having to negotiate, particularly in the online public space, and this is to see it in terms of the provision of a service. In the online public space and the world of Big Tech distinctions between platform and service, infrastructure and service are becoming more slippery and less clear. In the online public space, the provision of an online catalogue is also the provision of a search service and in such provision, individual archive services are never going to be able to compete with those being built by Big Tech. As McKean from the British Library explains *'Yeah, it's hard to find the balance without the financial backing it. If the industry might have something like Google it's going to be impossible to ever have an algorithm as sophisticated as that, the best case scenario is that in sort of ten, fifteen years time, you have what Google was ten, fifteens years ago. And researchers' expectations will have changed even more. So we are just playing catch up to some extent.'*

This perspective on the provision of access as the provision of a service was expressed and appeared, unsurprisingly given the point made in the previous paragraph, to be closely related to that of seeing it as the provision of a technical infrastructure. Interviewees who spoke in terms of one of these perspectives also spoke in terms of the other. For example, the Technical Lead on Project Alpha spoke of how: *'…the kind of development that government digital service and others are [expecting], so making sure that you are delivering digital services that work for users regardless of whether they've got a great network connection, regardless of whether they might be using an old device, regardless of whether they might be using assistive technology.'*

Then again Izzo was very clear in speaking of Legislation.gov.uk in similar terms: *'…but to answer that question we need to ask ourselves what type of service we are providing. Is it a service that…it's not an entertainment service right? It's an evidence-based type of platform, what we want is to report the law as it is published as it is. That is our purpose, that is the nature of the website and I think it will stay like that.'* Additionally, it seemed that it was from

this provision of a service perspective that questions started to rise about charging, the service being something that you charge for. Ordelman and Tweed both see this as an option when digitising material - the provision of a digitisation service, as they see it as something that users could contribute to instead of all costs falling on the institutions themselves. Then again Corti has had users willing to pay for the digitisation of collections and thinks archives could make more use out of the service provision: *'I think quite often archives don't make enough of the fact that users will quite often pay to get things done, that should always be exploited, really.'*

5.4 Navigating the online public space

This chapter has sought to elaborate and articulate further perspectives through which digital preservation practitioners are negotiating the provision of access, particularly in the online public space. It has therefore taken that space as its starting point and introduced, and drawn on perspectives from a different group, also operating in that space, in the form of data journalists. This group has been introduced through both the consultation of relevant literature and a number of interviews with those identifying as such.

As a result, a picture has been drawn of the online public space as to some extent a hostile environment for both groups. Some of the reasons why they find it hostile relate to the way in which certain patterns of behaviour and expectations seem to have been built into social interaction in the online public space which do not align well with both group's concerns with trust, truth and transparency. Other reasons come from the need for legal and ethical frameworks to develop fast to deal appropriately with this new space, which is not always happening. Both groups are becoming more proficient at operating in this space and both developing new skills and finding new ways of acting in order to help them to do that.

Looking through the frame of the online public space, an additional two perspectives were identified. These were, firstly that it was and is possible to look at and negotiate the provision of access to digital material in terms of the provision of access to a service and that it is possible to look at it in terms that also take into consideration the reception and subsequent use and interpretation of the digital material to which access is being given. The first of these was found to correlate quite closely with an earlier perspective that viewed the provision of access in terms of the provision of a technical infrastructure.

6 Conclusion

This research set out to explore the disconnect between what digital preservation practitioners are envisioning when talking about access to digital material and what is actually being made available in practice. To investigate this disconnect it was decided to review relevant literature as a base from which to understand the different perspectives and feelings of digital preservation practitioners and to enhance this understanding by interviewing participants working within this space or spaces adjacent to it.

The research consisted of three reflective research phases. One focusing on digital preservation practitioners, another on participants imagining access as computational access and the final one focusing on data journalists. The data gathered from literature and interviews and used within this work mainly comes from the digital preservation community that is active in North America, Australia and Europe. However, the research did suggest that whilst the themes arising do reference specific geographical difficulties, for example legislation or copyright, they could with caution be seen to be generally applicable.

It should be noted that this PhD was of a collaborative nature and therefore the chance to experiment with a number of concepts and ideas was possible. Further details on this can be found in Chapter 2, however the ways that these projects contributed to drawing the final conclusions, and especially the recommendations, will become apparent in this section. TNA was the partner in this work and even if they are not the main focus of research, most projects were run in parallel with them, and therefore, the recommendations are applicable to their practices but can also be applied to the wider digital preservation community.

The main research question addressed by this work was:

'How is digital preservation practice negotiating the provision of access in the digital environment?'

This question was first explored in Chapter 3 through a technological lens and an exploration of the evolving iterations of the internet. As the internet, or the online public space as it is referred to in this work, is seen as the most suitable place to make digital material available, it was deemed necessary to explore how online access had developed through these iterations. This exploration was difficult, as the history of the systems archival institutions have used to provide access is not well documented, but it became clear that the provision of access through the different iterations of the web has become more and more dependent on the technical infrastructure and the way in which this material is approached.

The infrastructure can no longer be seen as separate from the material, and the provision of access is reliant on the infrastructure and technology that is used. Technical professionals interviewed were quick to point this out and believe that the technology should become part of the core infrastructure of an organisation in order to guarantee it is both sustainable and interoperable. These individuals discussed a number of ways in which this could be accomplished including using an API-first approach and open-source tools when possible.

However, it did not take long to surface that even when viewing the provision of access through a technological lens, it could not be seen as purely a technological problem. Constraints around legislation and ethics were mentioned early on in the interviews with participants from digital preservation practice, highlighting how the provision of access was seen as consisting of a much wider set of tasks than just providing the correct technological infrastructure for this material to be made available on.

This led to a focus on the practice of digital preservation, which was explored in more detail in Chapter 4. This chapter considered the provision of access through the lens of digital preservation practice and reviewed how this practice had emerged and how the relationship between preservation and access had been balanced by it. When talking to participants about this relationship, it became clear that some of those participants identifying as digital preservation practitioners saw the importance of providing a sustainable and interoperable infrastructure but were nonetheless unsure what exactly this meant in practice.

This uncertainty around how to deliver such an infrastructure seemed to cause some friction and frustration on behalf of the more technically minded participants, as they did have a clear idea on what the digital infrastructure should look like but found it difficult to communicate this idea to their less technically minded colleagues. The general consensus of those interviewed seemed to be that the technical skills of digital preservation practitioners needed to increase; not in the sense that they all needed to become programmers themselves, but to a degree that would enable better communication with the technical individuals who are also operating in this space. Were this communication to be improved, digital preservation practitioners could leave the more technical aspects to the specialists and focus on their other concerns around digital material; which seemed to focus on providing adequate documentation to contextualise digital material and advocating for the importance of digital preservation.

Chapter 5 introduced the data journalists. These individuals were not preserving digital material, but were publishing digital content in the same online public space as the digital

preservation practitioners. They seemed to share similar core values with the digital preservation practitioners in that they regarded trust and provenance as two important aspects of their work. When talking to these individuals about their role, they also mentioned the impact of the digital shift and of thinking differently about digital material. They expressed a similar view on the fact that there was a need to work together with more technically minded people in the same space, and to learn how to communicate better with these individuals.

The online public space was explored in detail in Chapter 5. Here findings are discussed about how, to date, Big Tech has had a major/leading influence on and has created expectations around the provision of access in this space, to some extent making the space a hostile environment to operate in for both the data journalists and digital preservation practitioners. The analysis suggested that this was due to the fact that both data journalists and digital preservation practitioners are concerned with more than just delivering a service to users, but also with how the material to which they have provided access is being used and interpreted.

To summarise, this research has highlighted the complexity of making digital material available in the online public space but has also showcased how digital preservation practitioners may want to think about this material in the future in order to improve access. Firstly, this digital material cannot be seen as separate from the infrastructure anymore; access to digital material is dependent on it. However, there are individuals who can help with building this infrastructure. Secondly, the core values of digital preservation practitioners are important and should be maintained, especially as they concern the ethical and legislative considerations around access to digital material. In order to allow them to focus on such considerations, they must learn to communicate better with and to trust the technical professionals acting within the digital preservation space. Thirdly, the way the online public space has been shaped (e.g. mainly by Big Tech) may have influenced the expectations of users, but this does not necessarily mean that digital preservation practitioners should be making their material available in line with these expectations. The provision of access should not be seen merely as the provision of a service; taking into account the consequences of making this material available in the online space is also very important.

This research is a first step in exploring ideas and trends around the provision of access as this work has made it possible to look at the provision of access from a number of

perspectives. Therefore, it will make it easier for digital preservation practitioners to negotiate access in the future because it has unpicked a number of different perspectives through which they need to think about the provision of access; as the provision of an infrastructure, as the provision of adequate documentation, as the provision of a service, but also as a concern with acting legally and ethically with a view towards what the provision of this digital material will entail, and how people will impact and use it. Separating out these perspectives enables practitioners to act without the whole becoming a confusing mess; it rather provides a clearer view that helps them understand the perspectives they should use, and switch between and concentrate on these perspectives.

This is also where the element of negotiation comes into the work. The Oxford definition for the word negotiation states: *'formal discussion between people who are trying to reach an agreement'* (Oxford University Press 2022). Within this context and seeing this through different perspectives, it is possible for the digital preservation practitioners to see the tensions they will need to balance and negotiate when considering access to digital material. By splitting it into three manageable perspectives it is possible for the practitioners to address the issue of the provision of access with more balance and a better compromise.

Below a number of recommendations will be discussed based on the different perspectives which have been provided during this research. When looking at it from a technical infrastructure perspective, there is an emphasis on having the technology as a core part of the institution. Within the digital preservation community there is a move towards doing this, as demonstrated by both the literature and the interviewed participants. But an important part within this discussion that was stressed was the need for better communication between the digital preservation practitioners and the more technically minded individuals within this space.

There is a growing amount of literature and courses available to increase digital skills for practitioners (Underwood et al. 2018; Birkbeck 2022), but these are not necessarily focused on increasing communication and confidence in digital skills, but more on teaching individuals to do these technical implementations themselves. This is not necessarily what digital preservation practitioners should be aiming to do. It is true that digital skills should be enhanced but making it possible to identify and talk with technical individuals specialised in these skills is deemed more useful than trying to implement technical solutions themselves.

Further research into creating courses and resources to improve communication are of importance here. This was partly explored through running the Machine Learning Club (Bell

and Talboom 2022), with the caveat of being conducted on a very small scale and tailored towards a specific audience. However, this does highlight a number of recommendations from the technical perspective: increase the communication and confidence in technical skills. Hopefully this better understanding will then lead to technology becoming part of the core funding of an institution and, if possible, align with the technically minded individuals view on the best infrastructure, that currently seems to be an API-first approach.

A further perspective is on the provision of adequate documentation. Especially in light of viewing digital archival material as data this can be seen as something that needs to be further explored and therefore a further recommendation for organisations. A number of participants touched upon this during the interviews, yet there is not a lot happening in practice at this time. There is a move towards viewing this as important, even within fields that are more technically catered, such as the computer scientists who have learned a number of lessons from the archives and call their own practices *'the wild-west'* (Jo and Gebru 2020) highlighting their understanding of the meaningful contribution that digital preservation practitioners could make in this space. Noteworthy is the positive impact on the machine learning community in the form of a template for the documentation of datasets (Gebru et al. 2018).

Then there is the perspective of seeing access as a service, which is not looked at in detail, but was raised in the work that I was doing, this is discussed in detail in Chapter 5. But this is all talked about from a conceptual point of view. However, it would be of interest to see how this may look in practice, especially considering the monetisable value of digital material. This does not mean that everything needs to come with a cost, but it should be taken into consideration that if users want to compute over material as data on platforms provided by the organisations that this may come with additional costs. An example of this is the Google Colab platforms that provide users with a certain amount of computing power before charging for additional data runs. It is not being suggested that the same model should be decided upon by digital repositories, but this service model, and similar ones, could be explored in more detail while also taking into consideration the ethical implications of doing this.

The last perspective, and this is especially important when viewing this material from a wider context, is the ethical and legal consideration of this work. Individuals within the field push for the digital preservation practitioners to join in with the more technical debates on how technology should be moving forward (Johnson 2019; Kennedy 2019; Kilbride 2020).

Currently there is still a lack of bringing this into practice, which is understandable as it is novel, but there is certainly room for experimentation and exploration here.

To summarise, the following recommendations have been made:

- Ensuring the digital infrastructure is part of core funding
- Improving communication between technical individuals and digital preservation practitioners
- Exploring the documentation of digital material and seeing this as a valuable contribution from the digital preservation community
- Exploring the cost and use of a service model when making digital material accessible
- Taking into consideration the ethical and legal considerations of access into practice

Splitting the negotiation of access out in these perspectives has made it possible to provide helpful recommendations for all the different perspectives. This should therefore make it possible for digital preservation practitioners to move forward and work on the complex issue that is access to digital material.

The limitations of this work have been around focusing on countries who have been active within the digital preservation community for a number of decades, but this could be extended to countries that are starting to explore digital preservation. A large number of institutions that were interviewed during this work were limited by national legislation or policies that made access to this material difficult, this may be different for countries where this work is just starting. A lot of this project has been around creating a first step in making the negotiation around access to digital material possible for digital preservation practitioners, but a lot will still need to be worked out in practice. The collaborative nature of this PhD made it possible to combine research and work in practice, but as can be seen from the recommendations, a lot of work still needs to be done in this area.

References

Note on referencing style: *All references are provided in Chicago Manual of Style (17th edition) in the author-date configuration. There is a slight modification made to this referencing style to include the accessed date for webpages. This decision was made as webpages and web resources in general are a large part of this research. It is important to know when resources have been accessed, as they may have disappeared or have been modified. The choice was also made to include extra information on making it clear when a resource was an archived web resource.*

Abram, Stephen. 2007. 'Web 2.0, Library 2.0 and Librarian 2.0: Preparing for the 2.0 World'. In *Library and Information Services in Astronomy V (LISA V)*, edited by S. Ricketts, C. Birdie, and E. Isaksson, 377:161–67. Cambridge, MA: Astronomical Society of the Pacific.

AIM 25, Artefactual systems, Docuteam, Imagiz, Nothing Interactive, and Zazuko. 2018. 'AtoM 3 Proof of Concept Proposal'. AtoM Foundation Board of Directors.

AIMS Work Group. 2012. 'AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship'.

Alcantara, Chris, Kevin Schaul, Gerrit De Vynck, and Reed Albergotti. 2021. 'How Big Tech Got so Big: Hundreds of Acquisitions'. Washington Post. 2021. Accessed 16 November 2021. https://www.washingtonpost.com/technology/interactive/2021/amazon-apple-facebook-google-acquisitions/.

Amazon. 2020. 'Amazon Web Services (AWS) - Cloud Computing Services'. AWS. 2020. Accessed 19 August 2020. https://aws.amazon.com/.

American Library Association. 2008. 'Definitions of Digital Preservation'. American Library Association. Accessed 7 October 2021. http://www.ala.org/alcts/resources/preserv/defdigpres0408.

Ames, Sarah, and Lucy Havens. 2021. 'Exploring National Library of Scotland Datasets with Jupyter Notebooks': *IFLA Journal*, December. https://doi.org/10.1177/03400352211065484.

Anderson, Ian. 2008. 'Necessary but Not Sufficient: Modelling Online Archive Development in the UK'. *D-Lib Magazine* 14 (1/2). https://doi.org/10.1045/january2008-anderson.

Andresen, Herbjørn. 2019. 'A Discussion Frame for Explaining Records That Are Based on Algorithmic Output'. *Records Management Journal* 30 (2): 129–41. https://doi.org/10.1108/RMJ-04-2019-0019.

Appel, Rachel, Alison Clemens, Wendy Hagenmaier, and Jessica Meyerson. 2015. 'Born-Digital Access in Archival Repositories: Mapping the Current Landscape'. Preliminary Report.

ARIADNE. 2017. 'Final Report Summary - ARIADNE (Advanced Research Infrastructure for Archaeological Dataset Networking in Europe)'. Università degli Studi di Firenze. Accessed 23 June 2022. https://cordis.europa.eu/project/id/313193/reporting.

Arroyo-Ramírez, Elvia, Kelly Bolding, Danielle Butler, Alston Cobourn, Brian Dietz, Jessica Farrell, Alissa Helms, et al. 2020. 'Levels of Born-Digital Access'. Alexandria, VA: Digital Library Federation.

Bailey, Jefferson. 2018. 'Pseudodoxia Data: Our Ends Are as Obscure as Our Beginnings'. In *Always Already Computational: Collections as Data, Final Report*, edited by Thomas Padilla, Laurie Allen, Hannah Frost, Sarah Potvin, Elizabeth Russey Roke, and Stewart Varner, 102–3.

Bak, Greg. 2016. 'Trusted by Whom? TDRs, Standards Culture and the Nature of Trust'. *Archival Science* 16 (4): 373–402. http://dx.doi.org.libproxy.ucl.ac.uk/10.1007/s10502-015-9257-1.

Barr, Caelainn, Mona Chalabi, and Nick Evershed. 2019. 'A Decade of the DataBlog: "There's a Human Story behind Every Data Point"'. *The Guardian Datablog* (blog). 2019. Accessed 23 September 2020. https://www.theguardian.com/membership/datablog/2019/mar/23/a-decade-of-the-datablog-theres-a-human-story-behind-every-data-point.

Baxter, Terry D. 2011. 'Going to See the Elephant: Archives, Diversity, and the Social Web'. In *A Different Kind of Web: New Connections Between Archives and Our Users*, edited by Kate Theimer, 274–303. Chicago: Society of American Archivists.

Beavan, David, Fazl Barez, Mark Bell, John Fitzgerald, Eirini Goudarouli, Konrad Kollnig, Barbara McGillivray, et al. 2021. 'Discovering Topics and Trends in the UK Government Web Archive'. Data Study Group Final Report. London: Alan Turing Institute.

Bell, Mark, and Leontien Talboom. 2022. 'More than Just Algorithms: Machine Learning for Information Specialists'. In *The Rise of AI: Implications and Applications of Artificial Intelligence in Academic Libraries*, edited by S. Hervieux and A. Wheatley. Chicago: Association of College and Research Libraries.

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. 'The Semantic Web'. *Scientific American* 284 (5): 34–43.

Bettivia, Rhiannon S. 2016. 'The Power of Imaginary Users: Designated Communities in the OAIS Reference Model', Proceedings of 79th the Association for Information Science and Technology: Creating Knowledge, Enhancing Lives through Information & Technology, 53 (1): 38–46.

Birkbeck. 2022. 'Applied Data Science (Postgraduate Certificate)'. Birkbeck - University of London. 2022. Accessed 19 May 2022. https://www.bbk.ac.uk/study/2022/postgraduate/programmes/TPCCOMIP_C/0/applied-data-science-postgraduate-certificate.

Boczkowski, Pablo J., and Eugenia Mitchelstein. 2021. *The Digital Environment: How We Live, Learn, Work, and Play Now*. Cambridge, MA, USA: MIT Press.

Bourg, Chris. 2017. 'What Happens to Libraries and Librarians When Machines Can Read All the Books?' *Feral Librarian* (blog). 16 March 2017. Accessed 24 January 2019. https://chrisbourg.wordpress.com/2017/03/16/what-happens-to-libraries-and-librarians-when-machines-can-read-all-the-books/.

Brandtzaeg, Petter Bea, and Lüders Marika. 2018. 'Time Collapse in Social Media: Extending the Context Collapse'. *Social Media + Society* 4 (1): 1–10.

Braun, Virginia, and Victoria Clarke. 2006. 'Using Thematic Analysis in Psychology'. *Qualitative Research in Psychology* 3 (2): 77–101. http://dx.doi.org.libproxy.ucl.ac.uk/10.1191/1478088706qp063oa.

———. 2019. 'Reflecting on Reflexive Thematic Analysis'. *Qualitative Research in Sport, Exercise and Health* 11 (4): 589–97. https://doi.org/10.1080/2159676X.2019.1628806.

———. 2020. 'One Size Fits All? What Counts as Quality Practice in (Reflexive) Thematic Analysis?' *Qualitative Research in Psychology* 0 (0): 1–25. https://doi.org/10.1080/14780887.2020.1769238.

British Library. 2021a. 'British Library'. Flickr. 2021. Accessed 27 May 2021. https://www.flickr.com/photos/britishlibrary/.

———. 2021b. 'UK Web Archive'. British Library. 2021. Accessed 23 June 2021. https://www.bl.uk/collection-guides/uk-web-archive.

Bruseker, George, Nicola Carboni, and Anaïs Guillem. 2017. 'Cultural Heritage Data Management: The Role of Formal Ontology and CIDOC CRM'. In *Heritage and Archaeology in the Digital Age: Acquisition, Curation, and Dissemination of Spatial Cultural Heritage Data*, edited by Matthew L. Vincent, Victor Manuel López-

Menchero Bendicho, Marinos Ioannides, and Thomas E. Levy, 93–132. New York: Springer.

Bunn, Jenny. 2019. 'Looking Back at Archives on the Internet'. *ARC Magazine*, 2019.

———. 2020. 'Working in Contexts for Which Transparency Is Important: A Recordkeeping View of Explainable Artificial Intelligence (XAI)'. *Records Management Journal* 30 (2): 143–53. https://doi.org/10.1108/RMJ-08-2019-0038.

Burgess, Jamie Lynne. 2019. 'What's "Context Collapse"? Understanding It Can Mean a More Fulfilling Online Life'. *Rewire* (blog). 2019. Accessed 30 November 2020. https://www.rewire.org/context-collapse-online/.

Cambridge Dictionary. 2020. 'Computational'. In *Cambridge Dictionary*. Online. Accessed 27 May 2020. https://dictionary.cambridge.org/dictionary/english/computational.

Candela, Gustavo, María Dolores Sáez, MPilar Escobar Esteban, and Manuel Marco-Such. 2020. 'Reusing Digital Collections from GLAM Institutions'. *Journal of Information Science*, August, 0165551520950246. https://doi.org/10.1177/0165551520950246.

Candy, Linda. 2006. 'Practice Based Research: A Guide'. Sydney: Creativity & Cognition Studios.

CCSDS. 2002. 'Reference Model for an Open Archival Information Systems (OAIS)'. CCSDS 650.0-B-1. Washington: CCSDS.

———. 2012a. 'Reference Model for an Open Archival Information Systems (OAIS)'. CCSDS 650.0-M-1. Washington: CCSDS.

———. 2012b. 'Space Data and Information Transfer Systems -- Audit and Certification of Trustworthy Digital Repositories'. ISO 16363:2012. London: BSI Standards.

———. 2012c. 'Space Data and Information Transfer Systems - Open Archival Information System (OAIS) - Reference Model'. ISO 14721:2012. London: BSI Standards.

———. 2019. 'Reference Model for an Open Archival Information Systems (OAIS)'. CCSDS 650.0-P-3. Washington: CCSDS.

Center for Research Libraries. 2021. 'Certification and Assessment of Digital Repositories'. 2021. Accessed 1 December 2021. https://www.crl.edu/archiving-preservation/digital-archives/certification-assessment.

Charmaz, Kathy. 2006. *Constructing Grounded Theory: A Practice Guide through Qualitative Analysis*. Thousand Oaks, CA: Sage.

———. 2008. 'Constructionism and the Grounded Theory Method'. In *Handbook of Constructionist Research*, edited by J. A. Holstein and J. F. Gubrium, 397–412. New York: The Guildford Press.

Choak, C. 2012. 'Asking Questions: Interviews and Evaluations'. In *Research and Research Methods for Youth Practitioners*, edited by S. Bradford and F. Cullen, 90–112. London: Routledge.

CLARIAH. 2020. 'Media Suite User Documentation'. CLARIAH. Accessed 10 April 2020. https://mediasuite.clariah.nl/documentation/introduction.

Clarke, Victoria, and Virginia Braun. 2017. 'Thematic Analysis'. *The Journal of Positive Psychology* 12 (3): 297–98. https://doi.org/10.1080/17439760.2016.1262613.

CMoA. 2017. 'The Collections Data of the Carnegie Museum of Art in Pittsburgh, Pennsylvania'. Github. 2017. Accessed 2 June 2020. https://github.com/cmoa/collection.

Colavizza, Giovanni, Tobias Blanke, Charles Jeurgens, and Julia Noordegraaf. 2021. 'Archives and AI: An Overview of Current Debates and Future Perspectives'. *ArXiv:2105.01117 [Cs]*, May. Accessed 28 September 2021. http://arxiv.org/abs/2105.01117.

Coleman, Catherine Nicole. 2017. 'Artificial Intelligence and the Library of the Future, Revisited'. *Stanford Libraries* (blog). 3 November 2017. Accessed 24 January 2019. https://library.stanford.edu/blogs/digital-library-blog/2017/11/artificial-intelligence-and-library-future-revisited.

Cook, Terry. 1994. 'Electronic Records, Paper Minds: The Revolution in Information Management and Archives in the Post-Custodial and Post-Modernist Era'. *Archives and Manuscripts* 22 (2): 300–328.

Core Trust Seal. 2021. 'Core Certified Repositories'. 2021. Accessed 1 December 2021. https://www.coretrustseal.org/why-certification/certified-repositories/.

Corrado, Edward M., and Rachel Jaffe. 2014. 'Transforming and Enhancing Metadata for End User Discovery: A Case Study'. *JLIS.It* 5 (2): 33–48. http://dx.doi.org.libproxy.ucl.ac.uk/10.4403/jlis.it-10069.

Costa, Miguel, and Mário Silva. 2012. 'Evaluating Web Archive Search Systems'. In *Web Information Systems Engineering - WISE 2012*, edited by X.S. Wang, I. Cruz, A. Delis, and G. Huang. Vol. 7651. Lecture Notes in Computer Science. Accessed . https://doi.org/10.1007/978-3-642-35063-4_32.

CRL and OCLC. 2007. 'Trustworthy Repositories Audit & Certification: Criteria and Checklist'. Chicago: CRL & OCLC.

Davis, Jenny L, and Nathan Jurgenson. 2014. 'Context Collapse: Theorizing Context Collusions and Collisions'. *Information, Communication & Society* 17 (4): 476–85.

Davis, Susan. 2008. 'Electronic Records Planning in "Collecting" Repositories'. *The American Archivist* 71 (1): 167–89.

Digging into Data Challenge and Trans-Atlantic Platform. 2019. 'Digging into Data Challenge'. Digging into Data. 2019. Accessed 28 July 2020. https://diggingintodata.org/about.

Digital Heritage Network. 2019. 'National Digital Heritage Strategy'. Den Haag: Digital Heritage Network. Accessed 26 September 2019. https://www.netwerkdigitaalerfgoed.nl/wp-content/uploads/2018/10/20150608_Nationale_strategie_digitaal_erfgoed_Engels.pdf.

Digital Preservation Coalition. 2015. *Digital Preservation Handbook*. 2nd ed. Glasgow: Digital Preservation Coalition.

———. 2021. 'Novice to Know-How: Online Digital Preservation Training'. DPC Online. 2021. Accessed 1 December 2021. https://www.dpconline.org/digipres/train-your-staff/n2kh-online-training.

———. 2022a. 'Digital Preservation Coalition - Member List'. DPC Online. 2022. Accessed 3 March 2022. https://www.dpconline.org.

———. 2022b. 'Lowering the Barriers to Computational Access for Digital Archivists : A Launch Event'. Digital Preservation Coalition. 2022. Accessed 19 May 2022. https://www.dpconline.org/events/event-lowering-barriers-comp-access.

D'Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism*. Cambridge, MA: MIT Press.

DiNucci, Darcy. 1999. 'Fragmented Future'. *Print Magazine*, 1999.

Donnelly, Martin, Perla Innocenti, Andrew McHugh, and Raivo Ruusalepp. 2009. 'DRAMBORA Interactive User Guide'. Glasgow: DCC and DPE.

Doorn, Peter, and Heiko Tjalsma. 2007. 'Introduction: Archiving Research Data'. *Archival Science* 7 (1): 1–20. https://doi.org/10.1007/s10502-007-9054-6.

Drapeau, Martin. 2018. 'Our Friends CSV and JSON'. *Medium* (blog). 2018. Accessed 2 June 2020. https://medium.com/@martindrapeau/the-state-of-csv-and-json-d97d1486333.

Duff, Wendy M., Catherine A. Johnson, and Joan M. Cherry. 2013. 'Reaching out, Reaching in: A Preliminary Investigation into Archives' Use of Social Media in Canada'. *Archivaria*, no. 75: 77–96.

Edmond, Jennifer, and Vicky Garnett. 2015. 'APIs and Researchers: The Emperor's New Clothes?' *The International Journal of Digital Curation* 10 (1): 287–97. https://doi.org/10.2218/ijdc.v10i1.369.

Essen, Mette van. 2019. ‘Digital Preservation - Wat Is Het Niet (of Niet Alleen)’. In *Preserveren: Stappen Zetten in Een Nieuw Vakgebied*, edited by Margriet van Gorsel, Erika Hokke, Bart de Nil, and Marcel Ras, 135–41. Jaarboek 19. ’s-Gravenhage: Stichting Archiefpublicaties.

Etikan, Ilker. 2016. ‘Comparision of Snowball Sampling and Sequential Sampling Technique’. *Biometrics & Biostatistics International Journal* 3 (1). Accessed 13 May 2022. https://www.academia.edu/21810338/Comparision_of_Snowball_Sampling_and_Sequential_Sampling_Technique.

Europeana Foundation. 2020. ‘Our APIs’. Europeana Pro. 2020. Accessed 10 April 2020. https://pro.europeana.eu/page/apis.

———. 2022. ‘Our Mission’. Europeana Pro. 2022. Accessed 23 June 2022. https://pro.europeana.eu/about-us/mission.

Evans, Max J. 2007. ‘Archives of the People, by the People, for the People’. *The American Archivist* 70 (2): 387–400.

Everstijn, Carla. 2019. ‘The Digital Presence of Museums and the Implications for Collective Memory – MW19 | Boston’. In *MuseWeb Conference Proceedings*. Boston, MA: MuseWeb. Accessed 2 June 2022. https://mw19.mwconf.org/paper/the-digital-presence-of-museums-and-the-implications-for-collective-memory/index.html.

Gambini, Letizia. 2019. ‘A Decade of Working in Data Journalism: What Has Changed?’ Blog. *Europen Journalism Centre* (blog). 2019. Accessed 21 September 2020. https://medium.com/we-are-the-european-journalism-centre/a-decade-of-working-in-data-journalism-what-has-changed-8d950d99935e.

Garrett, John, and Donals Waters. 1996. ‘Preserving Digital Information: Report of the Taks Force on Archiving of Digital Information’. The Commision on Preservation and Access; The Research Libraries Group.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2018. ‘Datasheets for Datasets’, March. Accessed 25 June 2022. https://www.microsoft.com/en-us/research/publication/datasheets-for-datasets/.

Gerencser, James. 2011. ‘New Tools Equal New Opportunities: Using Social Media to Achieve Management Goals’. In *A Different Kind of Web: New Connections Between Archives and Our Users*, edited by Kate Theimer, 159–79. Chicago: Society of American Archivists.

GESIS – Leibniz-Institut für Sozialwissenschaften. 2019. 'Datenarchiv Für Empirische Sozialforschung'. Leibniz-Gemeinschaft. 2019. Accessed 30 May 2019. https://www.leibniz-gemeinschaft.de/infrastrukturen/archive/datenarchiv-fuer-empirische-sozialforschung/.

Giaretta, David. 2011. *Advanced Digital Preservation*. Berlin: Springer.

Gilliland, Anne J. 2016. 'Designing Expert Systems for Archival Evaluation and Processing of Computer-Mediated Communications: Frameworks and Methods'. In *Research in the Archival Multiverse*, edited by Anne J. Gilliland, Sue McKemmish, and Andrew J. Lau. Social Informatics. Clayton, Victori, Australia: Monash University Publishing.

GitHub. 2020. 'Build Software Better, Together'. GitHub. 2020. Accessed 19 August 2020. https://github.com.

Gollins, Tim. 2009. 'Parsimonious Preservation: Preventing Pointless Processes! (The Small Simple Steps That Take Digital Preservation a Long Way Forward)'. In *Online Information 2009 Proceedings*.

Gollins, Tim, and Emma Bayne. 2015. 'Finding Archived Records in a Digital Age'. In *Is Digital Different? How Information Creation, Capture, Preservation and Discovery Are Being Transformed*, edited by Michael Moss, Barbara Endicott-Popovsky, and Marc Dupuis. London: Facet Publishing.

Google. 2020. 'Welcome To Colaboratory'. Google Colab. 2020. Accessed 30 April 2020. https://colab.research.google.com/notebooks/intro.ipynb.

Goudarouli, Eirini. 2018. 'Computational Archival Science: Automating the Archive'. *The National Archives* (blog). 24 October 2018. Accessed 24 October 2018. https://blog.nationalarchives.gov.uk/blog/computational-archival-science-automating-archive/.

Goudarouli, Eirini, Anna Sexton, and John Sheridan. 2018. 'The Challenge of the Digital and the Future Archive: Through the Lens of The National Archives UK'. *Philosophy and Technology* 32 (1): 173–83.

Gray, Jonathan, Lucy Chambers, and Liliana Bounegru, eds. 2012. *The Data Journalism Handbook: How Journalists Can Use Data to Improve the News*. 1st edition. O'Reilly Media, Inc.

Guess, Andrew M., Brendan Nyhan, and Jason Reifler. 2020. 'Exposure to Untrustworthy Websites in the 2016 US Election'. *Nature Human Behaviour* 4 (5): 472–80. https://doi.org/10.1038/s41562-020-0833-x.

Guldi, Jo. 2018. ‘Critical Search: A Procedure for Guided Reading in Large-Scale Textual Corpora’. *Journal of Cultural Analytics* 3 (1): 11028. https://doi.org/10.22148/16.030.

HathiTrust. 2020. ‘HTRC Extracted Features Dataset’. HathiTrust. 2020. Accessed 13 July 2020. https://analytics.hathitrust.org/datasets.

Hawkins, Ashleigh. 2021. ‘Archives, Linked Data and the Digital Humanities: Increasing Access to Digitised and Born-Digital Archives via the Semantic Web’. *Archival Science*, December. https://doi.org/10.1007/s10502-021-09381-0.

Haynes, David, David Streatfield, Tanya Jowett, and Monica Blake. 1997. ‘Responsibility for Digital Archiving and Long Term Access to Digital Data’. London: British Library Research and Innovation Centre. Accessed 7 October 2021. http://www.ukoln.ac.uk/services/elib/papers/supporting/#blric.

Hedstrom, Margaret. 1998. ‘Digital Preservation: A Time Bomb for Digital Libraries’. *Computers and the Humanities* 31 (3): 189–202.

Higgins, Eliot. 2019. ‘Bellingcat and Beyond: The Future for Bellingcat and Online Open Source Investigation’. Keynote presented at the iPres 2019, Amsterdam, September 19.

Hillyard, Matthew. 2018. ‘Digital Archiving: The Seven Pillars of Metadata’. *The National Archives* (blog). 24 October 2018. Accessed 24 October 2018. https://blog.nationalarchives.gov.uk/blog/digital-archiving-seven-pillars-metadata/.

Hodge, Gail M. 2000. ‘Best Practices for Digital Archiving: An Information Life Cycle Approach’. *D-Lib Magazine* 6 (1). Accessed 7 October 2021. http://www.dlib.org/dlib/january00/01hodge.html#Garrett.

Hoffman, Chris. 2018. ‘What Is an API?’ *How-To Geek* (blog). 2018. Accessed 14 July 2020. https://www.howtogeek.com/343877/what-is-an-api/.

Holovaty, Adrian. 2006. ‘A Fundamental Way Newspaper Sites Need to Change’. Blog. *Adrian Holovaty - Writing* (blog). 2006. Accessed 21 September 2020. http://www.holovaty.com/writing/fundamental-change/.

Hooland, Seth van, and Ruben Verborgh. 2014. *Linked Data for Libraries, Archives and Museums : How to Clean, Link and Publish Your Metadata*. London: Facet Publishing. Accessed 19 October 2018.

Houston, Brant. 2015a. *Computer-Assisted Reporting: A Pratical Guide*. Fourth Edition. Abingdon: Routledge.

———. 2015b. ‘Fifty Years of Journalism and Data: A Brief History’. *Global Investigative Journalism Network*, 2015. Accessed 21 September 2020. https://gijn.org/2015/11/12/fifty-years-of-journalism-and-data-a-brief-history.

ICPSR. 2019. ‘ICPSR - Sharing Data to Advance Science’. ICPSR. 2019. Accessed 30 May 2019. https://www.icpsr.umich.edu/icpsrweb/.

International Image Interoperability Framework. 2022. ‘Gain Richer Access to the World’s Image and Audio/Visual Files’. IIIF. 2022. Accessed 15 June 2022. https://iiif.io/.

Jensen, Helle Strandgaard. 2020. ‘Digital Archival Literacy for (All) Historians’. *Media History*, 1–15. https://doi.org/10.1080/13688804.2020.1779047.

Jimerson, Randall C. 2007. ‘Archives for All: Professional Responsibility and Social Justice’. *The American Archivist* 70 (2): 252–81.

———. 2011. ‘Archives 101 in a 2.0 World: The Continuing Need for Parallel Systems’. In *A Different Kind of Web: New Connections Between Archives and Our Users*, edited by Kate Theimer, 75–101. Chicago: Society of American Archivists.

Jo, Eun Seo, and Timnit Gebru. 2020. ‘Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning’. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, January, 306–16. https://doi.org/10.1145/3351095.3372829.

Johnson, Sylvester A. 2019. ‘Technology Innovation and AI Ethics’. In *RLI 299: Ethics of Artificial Intelligence*, 14–27. Research Library Issues. Association of Research Libraries.

Jong, Winny de. 2021. ‘Data Journalism Tools - A Quick Guide to Find Tools to Analyse and Visualise Data’. Data Journalism Tools. 2021. Accessed 2 November 2021. https://datajournalism.tools/.

Kaplan, Andreas M., and Michael Haenlein. 2010. ‘Users of the World, Unite! The Challenges and Opportunities of Social Media’. *Business Horizons* 53 (1): 59–68.

Kärberg, Tarvo, and Koit Saarevet. 2016. ‘Transforming User Knowledge into Archival Knowledge’. *D-Lib Magazine* 22 (3/4).

Kayser-Bril, Nicolas. 2010. ‘Reasons to Cheer from Amsterdam’s Data-Driven Journalism Conference’. *Editors Blog - Journalism.Co.Uk* (blog). 2010. Accessed 21 September 2020. https://blogs.journalism.co.uk/2010/08/26/ddj-reasons-to-cheer-from-amsterdams-data-driven-journalism-conference/.

Kemman, Max, Martijn Kleppe, and Stef Scagliola. 2012. '"Just Google It"'. In *Proceedings of the Digital Humanities Congress 2012*, edited by Clare Mills, Michael Pidd, and Esther Ward. Sheffield: The Digital Humanities Institute.

Kennedy, Mary Lee. 2019. 'What Do Artificial Intelligence (AI) and Ethics of AI Mean in the Context of Research Libraries?' In *RLI 299: Ethics of Artificial Intelligence*, 3–13. Research Library Issues. Association of Research Libraries.

Kilbride, William. 2020. 'Nothing About Us Without Us'. *DPC Blog* (blog). 20 January 2020. Accessed 21 January 2020. https://dpconline.org/blog/nothing-about-us-without-us.

King, Gary. 2007. 'An Introduction to the Dataverse Network as an Infrastructure for Data Sharing'. *Sociological Methods and Research* 36: 173–99.

Koninklijke Bibliotheek. 2000. 'KB-Catalogus: Verfijnd Zoeken'. Archived Website. 2000. Accessed 22 July 2020. https://web.archive.org/web/20001021055315/http://www.kb.nl/kb/resources/frameset_kb.html?/kb/zoek/kbc_aqry.html.

———. 2001. 'Medieval Illuminated Manuscripts'. Archived Website. 2001. Accessed 2 July 2020. https://web.archive.org/web/20010624032059/http://www.kb.nl/kb/manuscripts/introduction/index.html.

Krause, Heather. 2019a. 'An Introduction to the Data Biography'. *We All Count* (blog). 2019. Accessed 14 August 2020. https://weallcount.com/2019/01/21/an-introduction-to-the-data-biography/.

———. 2019b. 'Fresh Apples to Old Oranges: A Case Study in Why Data Biographies Aren't Only Useful, They're Ethical'. *We All Count* (blog). 2019. Accessed 14 August 2020. https://weallcount.com/2019/01/24/fresh-apples-to-old-oranges-a-case-study-in-why-data-biographies-arent-only-useful-theyre-ethical/.

———. 2019c. 'Why We Need to Be Data Detectives'. *Medium* (blog). 2019. Accessed 14 August 2020. https://medium.com/@heatherkrause/why-we-need-to-be-data-detectives-7a55e289edf.

Kulesz, Octavio. 2017. 'Culture in the Digital Environment: Assessing Impact in Latin America and Spain'. Policy & Research. The United Nations Educational, Scientific and Cultural Organization. Accessed 9 May 2022. https://en.unesco.org/creativity/sites/creativity/files/dce-policyresearch-book2-en-web.pdf.

Kuny, Terry. 1998. 'The Digital Dark Ages? Challenges in the Preservation of Electronic Information'. In *Conference Programme and Proceedings*. Copenhagen, Denmark. Accessed 12 October 2021. https://www.semanticscholar.org/paper/The-digital-dark-ages-Challenges-in-the-of-Kuny/929b64974b18ff39b522e22c12365a1fca69824e.

Lanier, Jaron. 2018. 'Six Reasons Why Social Media Is a Bummer'. *The Observer*, 2018, sec. Technology. Accessed 16 November 2021. https://www.theguardian.com/technology/2018/may/27/jaron-lanier-six-reasons-why-social-media-is-a-bummer.

———. 2019. *Ten Arguments For Deleting Your Social Media Accounts Right Now*. London: Vintage.

Lankes, R. David. 2019. 'Decoding AI and Libraries'. *R. David Lankes* (blog). 3 July 2019. Accessed 24 January 2020. https://davidlankes.org/decoding-ai-and-libraries/.

Lavoie, Brian. 2014. 'The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition)'. DPC Technology Watch Report. Glasgow: Digital Preservation Coalition. https://doi.org/10.7207/twr14-02.

Lee, C.A. 2010. 'Open Archival Information System (OAIS) Reference Model'. In *Encyclopedia of Library and Information Sciences, Third Edition*, edited by M. J. Bates and M. N. Maack, Third. Boca Raton: CRC Press.

Levine, Alexandra S. 2021. 'Is Twitter Angling to Become Big Tech?' POLITICO. 2021. Accessed 16 November 2021. https://politi.co/3ksS4X6.

Li, Chunqiu, and Shigeo Sugimoto. 2018. 'Provenance Description of Metadata Application Profiles for Long-Term Maintenance of Metadata Schemas'. *Journal of Documentation* 74 (1): 36–61.

Library Carpentry. 2022. 'Library Carpentry - Software and Data Skills for People Working in Library- and Information-Related Roles'. Library Carpentry. 2022. Accessed 19 May 2022. https://librarycarpentry.org/.

Liew, Chern Li. 2016. 'Social Metadata and Public-Contributed Contents in Memory Institutions: "Crowd Voice" Versus "Authenticated Heritage"?' *Preservation, Digital Technology & Culture* 45 (3): 122–33. http://dx.doi.org.libproxy.ucl.ac.uk/10.1515/pdtc-2016-0017.

Liew, Chern Li, Vanessa King, and Gillian Oliver. 2015. 'Social Media in Archives and Libraries: A Snapshot of Planning, Evaluation, and Preservation Decisions'. *Preservation,*

Digital Technology & Culture 44 (1): 3–11. http://dx.doi.org.libproxy.ucl.ac.uk/10.1515/pdtc-2014-0023.

Locher, Anita E. 2016. 'Starting Points for Lowering the Barrier to Spatial Data Preservation'. *Journal of Map & Geography Libraries* 12 (1): 28–51. http://dx.doi.org/10.1080/15420353.2015.1080781.

———. 2019. 'Characterizing Potential User Groups for Versioned Geodata'. In *Service-Oriented Mapping. Changing Paradigm in Map Production and Geoinformation Management*, edited by J. Döllner, M. Jobst, and P. Schmitz, 417–34. Lecture Notes in Geoinformation and Cartography. New York: Springer.

Lohndorf, Jillian. 2022. 'Known Web Archiving Challenges'. Archive-It Help Center. 2022. Accessed 6 June 2022. https://support.archive-it.org/hc/en-us/articles/209637043-Known-Web-Archiving-Challenges.

Mahey, Mahendra, Aisha Al-Abdulla, Sarah Ames, Paula Bray, Gustavo Candela, Sally Chambers, Caleb Derven, et al. 2019. *Open a GLAM Lab*. QU Press. Accessed 30 September 2021. http://qspace.qu.edu.qa/handle/10576/12115.

Marciano, Richard, Victoria Lemieux, Mark Hedges, Maria Esteva, William Underwood, Michael Kurtz, and Mark Conrad. 2018. 'Archival Records and Training in the Age of Big Data'. In *Re-Envisioning the MLS: Perspectives on the Future of Library and Information Science Education*, edited by Johnna Percell, Lindsay C. Sarin, Paul T. Jaeger, and John Carlo Bertot, 44B:179–99. Advances in Librarianship. Emerald Publishing Limited. https://doi.org/10.1108/S0065-28302018000044B010.

Marwick, Alice E., and danah boyd. 2010. 'I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience'. *New Media & Society* 13 (1): 114–33. https://doi.org/10.1177/1461444810365313.

Mayer-Schonberger, Viktor, and Thomas Ramge. 2018. 'A Big Choice for Big Tech: Share Data or Suffer the Consequences World War Web'. *Foreign Affairs* 97 (5): 48–54.

McDonough, Jerome P. 2012. '"Knee-Deep in the Data": Practical Problems in Applying the OAIS Reference Model to the Preservation of Computer Games'. In *45th Hawaii International Conference on System Sciences*, 1625–34. Hawaii: IEEE Computer Society.

McGovern, Nancy. 2016. 'Current Status of Trustworthy Systems'. In *Building Trustworthy Digital Repositories: Theory and Implementation*, edited by Philip C. Bantin, 325–36. London: Rowman & Littlefield.

Melgar-Estrada, Liliana, Marijn Koolen, Kaspar Beelen, Hugo Huurdeman, Mari Wigham, Carlos Martinez-Ortiz, Jaap Blom, and Roeland Ordelman. 2019. 'The CLARIAH Media Suite: A Hybrid Approach to System Design in the Humanities'. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, 373–77. CHIIR '19. Glasgow, Scotland UK: ACM. https://doi.org/10.1145/3295750.3298918.

Meroño-Peñuela, Albert, Ashkan Ashkpour, Valentijn Gilissen, Jan Jonker, Tom Vreugdenhil, and Peter Doorn. 2018. 'Improving Access to the Dutch Historical Censuses with Linked Open Data'. *Research Data Journal for the Humanities and Social Sciences* 3 (1): 1–14.

Meyrowith, Joshua. 1985. *No Sense of Place: The Impact of Electronic Media on Social Behavior*. Oxford: Oxford University Press.

MoMA. 2020. 'The Museum of Modern Art (MoMA) Collection'. Github. 2020. Accessed 31 March 2020. https://github.com/MuseumofModernArt/collection.

Mordell, Devon. 2019. 'Critical Questions for Archives as (Big) Data'. *Archivaria* 87: 140–61.

Moss, Michael, David Thomas, and Tim Gollins. 2018. 'The Reconfiguration of the Archive as Data to Be Mined'. *Archivaria*, November, 118–51.

Muehlberger, Guenter, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, et al. 2019. 'Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study'. *Journal of Documentation* 75 (5): 954–76. https://doi.org/10.1108/JD-07-2018-0114.

Murambiwa, Ivan, and Patrick Ngulube. 2011. 'Measuring Access to Public Archives and Developing an Access Index: Experiences of the National Archives of Zimbabwe'. *ESARBICA Journal* 30: 83–101.

National Archives and Records Administration. 2000. 'The Book Catalog: Search the Library Online Public Access Catalog'. Archived Website. 2000. Accessed 22 July 2020. https://web.archive.org/web/20000817182516/http://www.nara.gov/alic/opac.html.

National Library of Scotland. 2020. 'Jupyter Notebooks'. National Library of Scotland. 2020. Accessed 27 May 2021. https://data.nls.uk/tools/jupyter-notebooks/.

Nestor. 2019. 'Nestor Seal for Trustworthy Digital Archives'. Nestor. 2019. Accessed 30 May 2019. http://www.dnb.de/Subsites/nestor/EN/Siegel/siegel.html.

Netwerk Digitaal Erfgoed. 2019. 'Erfgoed Digitaal Voor Allemaal: Intensivering van de Dienstverlening En de Inclusiviteit van Het Netwerk Digitaal Erfgoed 2019-2020'. Den Haag: Netwerk Digital Erfgoed.

Newman, James. 2018. 'The Game Inspector: A Case Study in Gameplay Preservation'. *Kinephanos: Journal of Media Studies and Popular Culture*, 120–48.

Nicholas, David, and David Clark. 2015. 'Finding Stuff'. In *Is Digital Different? How Information Creation, Capture, Preservation and Discovery Are Being Transformed*, edited by Michael Moss, Barbara Endicott-Popovsky, and Marc Dupuis, 19–34. London: Facet Publishing.

Nicholson, Bob. 2013. 'The Digital Turn: Exploring the Methodological Possibilities of Digital Newspaper Archives'. *Media History* 19 (1): 59–73. https://doi.org/10.1080/13688804.2012.752963.

Niu, Jinfang. 2016. 'Linked Data for Archives'. *Archivaria*, no. 82: 83–110.

Odell, Jenny. 2019. *How to Do Nothing: Resisting the Attention Economy*. Brooklyn, New York: Melville House.

O'Reilly, Tim. 2005. 'What Is Web 2.0 - Design Patterns and Business Models for the Next Generation of Software'. *O'Reilly* (blog). 2005. Accessed 13 June 2019. https://www.oreilly.com/pub/a//web2/archive/what-is-web-20.html.

Oxford University Press. 2022. 'Negotiation'. In *Oxford Dictionary*. Oxford: Oxford University Press. Accessed 8 December 2022. https://www.oxfordlearnersdictionaries.com/definition/english/negotiation?q=negotiation.

Padilla, Thomas, Laurie Allen, Hannah Frost, Sarah Potvin, Elizabeth Russey Roke, and Stewart Varner. 2018. 'Always Already Computational: Collections as Data'. Final Report.

Palmer, Joy. 2009. 'Archives 2.0: If We Build It, Will They Come?' *Ariadne*, no. 60. Accessed . http://www.ariadne.ac.uk/issue/60/palmer/.

Palmer, Joy, and Jane Stevenson. 2011. 'Something Worth Sitting Still For? Some Implications of Web 2.0 for Outreach'. In *A Different Kind of Web: New Connections Between Archives and Our Users*, edited by Kate Theimer. Chicago: Society of American Archivists.

Pearce-Moses, Richard. 2007. 'Janus in Cyberspace: Archives on the Threshold of the Digital Era'. *The American Archivist* 70 (1): 13–22.

Petram, Lodewijk, Elvin Dechesne, and Gijsbert Kruithof. 2018. 'Person Name Vocabulary'. The Hague: Huygens Institute for the History of the Netherlands. Accessed . https://lodewijkpetram.nl/vocab/pnv/doc/.

Pickard, Alison Jane. 2013. *Research Methods in Information*. Second edition. London: Facet. Accessed 14 December 2021. http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=nlebk&AN=1560617&site=ehost-live&scope=site.

Preservica. 2018. 'Protect and Secure Your Digital Information for Decades to Come'. Preservica. 19 November 2018. Accessed 19 November 2018. https://preservica.com/digital-archive-software.

Primary Trustworthy Digital Repository Authorisation Body. 2021. 'Certified Clients'. ISO 16363. 2021. Accessed 1 December 2021. http://www.iso16363.org/iso-certification/certified-clients/.

Project Jupyter. 2020. 'Project Jupyter'. 2020. Accessed 19 August 2020. https://www.jupyter.org.

Public Record Office. 2001. 'Public Record Office Catalogues: Procat'. Archived Website. 2001. Accessed 22 July 2020. https://web.archive.org/web/20010413163551/http://www.pro.gov.uk/catalogues/procat.htm.

Putnam, Lara. 2016. 'The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast'. *The American Historical Review* 121 (2): 377–402.

Rahimi, Alireza, Mohammad R. Soleymani, Alireza Hashemian, Mohammad R. Hashemian, and Azra Daei. 2018. 'Evaluating Digital Libraries: A Systematised Review'. *Health Information & Libraries Journal* 35 (3): 180–91. https://doi.org/10.1111/hir.12231.

Reynolds, Jonh F. 1998. 'Do Historians Count Anymore? The Status of Quantitative Methods in History, 1975-1995'. *Historical Methods* 31 (4): 141–48.

Richards, Lyn. 1999. *Using NVIVO in Qualitative Research*. SAGE.

Ridley, Michael. 2019. 'Explainable Artificial Intelligence'. In *RLI 299: Ethics of Artificial Intelligence*, 28–46. Research Library Issues. Association of Research Libraries.

RLG and OCLC. 2002. 'Trusted Digital Repositoires: Attributes and Responsibilities'. Mountain View, California: RLG. Accessed . https://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf.

Romein, C. Annemieke, Max Kemman, Julie M. Birkholz, James Baker, Michel De Gruijter, Albert Meroño-Peñuela, Thorsten Ries, Ruben Ros, and Stefania Scagliola. 2020. 'State of the Field: Digital History'. *History* 105 (365): 291–312. https://doi.org/10.1111/1468-229X.12969.

Ruest, Nick, Jimmy Lin, Ian Milligan, and Samantha Fritz. 2020. 'The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives'. *ArXiv:2001.05399*.

Saldaña, Johnny. 2009. *The Coding Manual for Qualitative Researchers*. London: SAGE Publications.

Sexton, Anna, Elizabeth Shepherd, Oliver Duke-Williams, and Alexandra Eveleigh. 2017. 'A Balance of Trust in the Use of Government Administrative Data'. *Archival Science* 17 (4): 305–30.

Sheridan, John. 2019. 'Keynote Speaker'. Presented at the Let's Get Digital, The National Archives, January 16.

Sherratt, Tim. 2020. 'Welcome to the Wonderful World of GLAM Data!' GLAM Workbench. 2020. Accessed 30 April 2020. https://glam-workbench.github.io/.

———. 2021. 'Web Archives'. GLAM Workbench. 2021. Accessed 28 June 2021. https://glam-workbench.net/web-archives/.

Shivalingaiah, D., and Umesha Naik. 2008. 'Comparative Study of Web 1.0, Web 2.0 and Web 3.0', February. Accessed 2 June 2022. http://ir.inflibnet.ac.in:8080/ir/handle/1944/1285.

Sierman, Barbara. 2019. 'OAIS: A restriction or a Change? [OAIS: Keurslijf of Kans?]'. In *Preserveren: Stappen zetten in een nieuw vakgebied*, edited by Margriet van Gorsel, Erika Hokke, Bart de Nil, and Marcel Ras, 36–44. Jaarboek 19. 's-Gravenhage: Stichting Archiefpublicaties.

Sierman, Barbara, and Marcel Ras. 2019. 'Certificering van Digitale Archieven in Nederland'. In *Preserveren: Stappen Zetten in Een Nieuw Vakgebied*, edited by Margriet van Gorsel, Erika Hokke, Bart de Nil, and Marcel Ras, 65–72. Jaarboek 19. 's-Gravenhage: Stichting Archiefpublicaties.

Srnicek, Nick. 2017. *Platform Capitalism*. Cambridge: Polity Press.

Stanford University's Department of Special Collections & University Archives. 2021. 'EPADD Installation and User Guide'. Accessed . https://docs.google.com/document/d/1CVIpWK5FNs5KWVHgvtWTa7u0tZjUrFrBHq6_6ZJVfEA/edit.

Stewart, Kelly, and Stefana Breitweiser. 2019. 'SCOPE: A Digital Archives Access Interface'. *The Code4Lib Journal*, no. 43 (February). Accessed 9 December 2021. https://journal.code4lib.org/articles/14283.

Storrar, Tom, and Leontien Talboom. 2019. 'Network Analysis of the UK Government Web Archive'. The National Archives Blog. 2019. Accessed 14 October 2019. https://blog.nationalarchives.gov.uk/network-analysis-of-the-uk-government-web-archive/.

Suber, Peter. 2015. 'Open Access Overview - Focusing on Open Access to Peer-Reviewed Research Articles and Their Preprints'. Legacy Page. Earlham College. 2015. Accessed 9 December 2021. http://bit.ly/oa-overview.

Talboom, Leontien. 2022. 'First Steps to a Guide for Computational Access to Digital Repositories'. *Software Sustainability Institute* (blog). 2022. Accessed 19 May 2022. https://www.software.ac.uk/blog/2022-05-04-first-steps-guide-computational-access-digital-repositories.

Talboom, Leontien and Digital Preservation Coalition. 2022. 'Computational Access: A Beginner's Guide for Digital Preservation Practitioners'. Digital Preservation Coalition. Accessed 4 November 2022. https://www.dpconline.org/digipres/implement-digipres/computational-access-guide.

Talboom, Leontien, and David Underdown. 2019. '"Access Is What We Are Preserving": But for Whom?' *DPC Blog* (blog). 2019. Accessed 14 October 2019. https://www.dpconline.org/blog/access-what-we-are-preserving.

Tasovac, Toma, Adrien Barbaresi, Thibault Clérice, Jennifer Edmond, Natalia Ermolaev, Vicky Garnett, and Clifford Wulfman. 2016. 'APIs in Digital Humanities: The Infrastructural Turn'. In , 93–96. Cracovie, Poland: HAL.

Terranova, Tiziana. 2000. 'Free Labor: Producing Culture for the Digital Economy'. *Social Text* 18 (June). https://doi.org/10.1215/01642472-18-2_63-33.

The Alan Turing Institute. 2021. 'Discussion Group Humanities & Data Science @Turing'. Notes. HackMD. 2021. Accessed 6 October 2021. https://hackmd.io/@turing-hds/DiscussionGroup.

The British Library. 2017. 'Sustaining The Value: The British Library Digital Preservation Strategy 2017-2020'. London: The British Library.

The Dataverse Project. 2022. ‘The Dataverse Project - Open Source Research Data Repository Software’. Dataverse.Org. 2022. Accessed 5 October 2022. https://dataverse.org/home.

The Economist. 2018. ‘The Tech Giants Are Still in Rude Health’. *The Economist*, 4 August 2018. Accessed 16 November 2021. https://www.economist.com/business/2018/08/04/the-tech-giants-are-still-in-rude-health.

The European Commission. 2019. ‘A Europe Fit for the Digital Age: Empowering People with a New Generation of Technologies’. 2019. Accessed 8 May 2022. https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age_en.

The National Archives. 2016. ‘RiC Feedback from the UK National Archives’. London: The National Archives.

———. 2017a. ‘Digital Cataloguing Practices at The National Archives’. London: The National Archives. Accessed . http://www.nationalarchives.gov.uk/documents/digital-cataloguing-practices-march-2017.pdf.

———. 2017b. ‘Digital Strategy 2017-2019’. London: The National Archives.

———. 2017c. ‘Legislation API Service Guide’. London: The National Archives.

———. 2019. ‘Plugged in, Powered up - A Digital Capacity Building Strategy for Archives’. London: The National Archives.

———. 2021a. ‘Discovery for Developers: About the Application Programming Interface (API)’. Text. Website Help and Terms of Use. 2021. Accessed 25 June 2021. http://www.nationalarchives.gov.uk/help/discovery-for-developers-about-the-application-programming-interface-api/.

———. 2021b. ‘MT 205’. Catalogue Entry. Discovery. 2021. Accessed 25 November 2021. https://discovery.nationalarchives.gov.uk/browse/r/h/C16488.

The Programming Historian. 2020. ‘About the Programming Historian’. The Programming Historian. 2020. Accessed 10 April 2020. https://programminghistorian.org/en/about.

The Royal Society. 2022. ‘The Online Information Environment: Understanding How the Internet Shapes People’s Engagement with Scientific Information’. London: The Royal Society. Accessed 18 February 2022. https://royalsociety.org/-/media/policy/projects/online-information-environment/the-online-information-environment.pdf.

The Steinmetz Archive. 1989. ‘Steinmetz Archive: Dutch Social Science Data-Archive’. *Historical Social Research / Historische Sozialforschung* 14 (1 (49)): 118–21.

Theimer, Kate. 2011. ‘What Is the Meaning of Archives 2.0?’ *The American Archivist* 74 (1): 58–68.

———. 2018. ‘It’s the End of the Archival Profession as We Know It, and I Feel Fine’. *Archival Futures*. https://doi.org/10.29085/9781783302192.002.

Thylstrup, Nanna Bonde. 2019. *The Politics of Mass Digitization*. https://doi.org/10.7551/mitpress/11404.001.0001.

Towards A National Collection. 2021. ‘Towards a National Collection | Collections United’. Towards A National Collection. 2021. Accessed 2 December 2021. https://www.nationalcollection.org.uk/.

Townsend, Robert B. 2011. ‘Old Divisions, New Opportunities: Historians and Other Users Working with and in Archives’. In *A Different Kind of Web: New Connections Between Archives and Our Users*, edited by Kate Theimer, 213–32. Chicago: Society of American Archivists.

Transport for London. 2021. ‘About the Catalogue’. Corporate Archives - TfL. 2021. Accessed 25 November 2021. http://www.tflcorporatearchivescatalogue.co.uk/CalmViewA/Aboutcatalogue.aspx.

UK Data Service. 2022. ‘Data Catalogue’. UK Data Service. 2022. Accessed 3 March 2022. https://beta.ukdataservice.ac.uk/datacatalogue/.

Underdown, David. 2018. ‘Using the Discovery API to Analyse Catalogue Data’. *The National Archives Blog*, 2018. Accessed 5 November 2018. https://blog.nationalarchives.gov.uk/blog/using-the-discovery-api/.

Underwood, William, David Weintrop, Michael Kurtz, and Richard Marciano. 2018. ‘Introducing Computational Thinking into Archival Science Education’. In *2018 IEEE International Conference on Big Data (Big Data)*, 2761–65. https://doi.org/10.1109/BigData.2018.8622511.

Vermaaten, Sally, Brian Lavoie, and Priscilla Caplan. 2012. ‘Identifying Threats to Succesful Digital Preservation: The SPOT Model for Risk Assessment’, D-Lib Magazine, 18 (9/10).

Wardley, Simon. 2017. ‘Future *is* Predictable’. First draft book. Accessed . http://www.wardleymaps.com/uploads/9/5/9/6/9596026/future-is-predictable-v12.pdf.

Wellcome Trust. 2020. 'Make Something - Use Open APIs and Datasets to Make Something New with Our Collections'. Wellcome Collection Developers. 2020. Accessed 14 July 2020. https://developers.wellcomecollection.org/.

Whitelaw, Mitchell. 2015. 'Generous Interfaces for Digital Cultural Collections'. *Digital Humanities Quarterly*. Accessed 28 May 2021. https://openresearch-repository.anu.edu.au/handle/1885/153515.

Wigham, Mari, Liliana Melgar Estrada, and Roeland J. F. Ordelman. 2019. 'Jupyter Notebooks for Generous Archive Interfaces'. In *2018 IEEE International Conference on Big Data (Big Data)*. https://doi.org/10.1109/BigData.2018.8622203.

Wiltshire, Gareth, and Noora Ronkainen. 2021. 'A Realist Approach to Thematic Analysis: Making Sense of Qualitative Data through Experiential, Inferential and Dispositional Themes'. *Journal of Critical Realism* 20 (2): 159–80. https://doi.org/10.1080/14767430.2021.1894909.

Winters, Jane, and Andrew Prescott. 2019. 'Negotiating the Born-Digital: A Problem of Search'. *Archives and Manuscripts* 47 (3): 391–403.

Yakel, Elizabeth. 2011. 'Balancing Archival Authority with Encouraging Authentic Voices to Engage with Records'. In *A Different Kind of Web: New Connections Between Archives and Our Users*, edited by Kate Theimer, 75–101. Chicago: Society of American Archivists.

Yeo, Geoffrey. 2013. 'Trust and Context in Cyberspace'. *Archives and Records* 34 (2): 214–34. https://doi.org/10.1080/23257962.2013.825207.

Yoon, Ayoung. 2014. 'End Users' Trust in Data Repositories: Definition and Influences on Trust Development'. *Archival Science* 14 (1): 17–34. http://dx.doi.org.libproxy.ucl.ac.uk/10.1007/s10502-013-9207-8.

Appendices

Appendix 1 – Topic Guides and Interview Questions

Topic Guide – Phase 1

The interviews in this research will be of a semi-structured nature. The main research question to be addressed is 'Why are digital preservation practitioners struggling to make born-digital data accessible'. The focus of this research will be on the struggle that these digital preservation practitioners are facing and how the are currently solving this problem. As it is unclear at the moment how this struggle has formed, it would not be useful to address this study with structured interview questions, therefore a semi-structured approach was chosen.

Below a list of topics that will be covered during these interviews are summarised. However, the interviews will not be limited to these topics and could potentially cover other topics arising during the interviews. This topic guide is a draft and may change over the course of the study.

What does accessibility mean for the practitioner?

Preserving analogue records was enough to ensure access and enable use to these records. However, this is different for born-digital data, where there are more elements that limit the use and accessibility to this data. It would be interesting to hear what the practitioners view is on access and what it means to make data accessible.

Methods to make born-digital data accessible

These will cover the current techniques and methods that digital preservation practitioners use to make the files accessible. But also, projects that the practitioners are working on and their future ideas for accessibility to born-digital data.

Users and designated communities

This topic will cover the user and how the digital preservation practitioners facilitate to their users. It would be interesting to know how they gain insights into their users and what they use this collected data for.

Designated community is a term closely related to user and is introduced in the OAIS model, which is a model used for digital preservation. In many models, checklists and certifications it is stated that a designated community is needed to run a digital archive. It would be interesting to hear what designated community mean for the participant and if their organisation has a designated community. This question is of specific interesting for

organisation that have a mandatory commitment to making their data available to a large audience, such as the general public.

Contextualising born-digital material

This topic will cover how the digital preservation practitioners contextualize their born-digital data. This will cover terms such as linked data, metadata and metadata standards. It would be good to know how the practitioner is contextualizing their born-digital data, especially in relation to how it was done for analogue data (If applicable to the organisation).

Automating part of the archiving process

As Machine Learning and Artificial Intelligence has grown in popularity in the last 10 years, archives are also starting to adapt these techniques. It would be interested to understand the archivist opinion on the use of these methods and tools and how they could see them as a beneficial attribute or as a drawback with the archival sector.

Responsibility to make born-digital data accessible

Preservation and accessibility have been terms that have been closely related, however, born-digital data makes the link between these two processes weaker, as preserving a born-digital file does not grant immediate access to it. It would be interesting to see how far the digital preservation practitioner goes to make this data accessible and what their thought is on who the responsible body is to make this data accessible.

Trust and certification

This topic will cover the methods that the digital preservation practitioner uses to gain trust from their users and what this trust means for the practitioner themselves.

There are a number of certifications out there for digital preservation practitioners to accredit their archives and show how they follow certain procedures. Getting an overview from the digital preservation practitioner on if they have certification or had certification in the past would be useful. It would also be useful to know why or why not they have decided to gain certification and what their main view is on certifications.

Interview Questions – Phase 1

What is your role in this organisation?

How long have you been working for this organisation?

<Ingest>

What data do you collect?

Where does the data come from?

How is the data sent to you?

What processing is done by the depositors?

Do you have a role in the selection process?

<Preservation>

How is this data processed? -> *Talk through step by step, so don't ask them what software they use, but ask them what the first step is, second step is, etc. Do not give examples, because they may focus on only the examples that I am giving.*

If it was possible, would you change anything about your current processing techniques and why?

How has the processing of data changed over the last few years/decade?

Have you seen changes to the processing of the data over the last few years?

Where do you store the data?

Is any of the processing automated or done by a machine?

Are there any constraints in the processing method?

Is any of the process outsourced to other institutions or companies?

Does anyone else have control over any of the processing stages?

<Accessibility>

How is this data made presentable?

Who uses your data?

How do you make users aware of your data?

In an ideal world, how would you make data accessible to your user group?

Do you restrict access to data?

How do you ensure trust in your organisation?

Why do you think users trust you?

How do you engage with your audience?

Topic Guide – Phase 2

This guide will give a brief overview of the topics likely to be discussed during the interview. As the interview is of a semi-structured nature, the interview may not be limited to these topics and could potentially cover other topics arising during the interview.

Facilitating the user to compute over the material

As you have been approached because you have been able to facilitate your users in accessing your material as data, it would be useful to get an understanding of the framework that has been put into place to facilitate this and why the decision was made to create this framework. This topic will also focus on how this framework is maintained and any other future plans for the framework.

The material made accessible

The framework will make it possible for the user to access certain or all material within your organisation. This topic will cover the decisions made when deciding what would be made available, but also what security measures have been put into place to ensure the material is used in a safe and responsible way. This topic will also cover the environment where the user is able to manipulate the material, is this within a restricted platform set up by you or is the user able to download the material to a local environment?

The users of the framework

As this framework will have certain users, it would be useful to discuss these in more detail, including the type of feedback that these users provide. Another interest is how much guidance is offered to the user when accessing this material through the framework. Is it presumed that the user should know how to navigate this or is there extra guidance available to help the user along?

Interview Questions – Phase 2

What is your role in this organisation?

How long have you been working for this organisation?

<The material>

What type of data do you provide access to? Is this data structured or unstructured?

Do you provide access to all your material or partial? And why?

How does versioning work for your data? Is this data static or changing?

What format is the data made available in and why?

What security measures do you have in place to ensure no misconduct of the data?

Do you have a take-down policy?

Are users able to download your material or is it only accessible through the framework?

<The framework>

What type of framework do you have in place for your users to access your material? And why?

Is everyone able to access this framework?

Who maintains this framework? And how do you ensure the framework is sustainable?

Would you change anything about the current framework?

Are there future plans in expanding or replacing the framework?

<Users>

Who uses this framework and for what purpose?

Do you collect feedback from your users? And if yes, what type of feedback do you receive?

Is there support for users or is there an assumption that they should know how to navigate this framework?

Topic Guide – Phase 3

This guide will give a brief overview of the topics likely to be discussed during the interview. As the interview is of a semi-structured nature, the interview may not be limited to these topics and could potentially cover other topics arising during the interview.

Documentation of data

Data journalists seem to take a slightly different approach to documenting the datasets that they use for their work. It would be interesting to hear how you document this dataset. This topic will also focus on how trust is maintained when providing access to these datasets, and how bias is recorded.

Making the material accessible

In the last few years, the interfaces and interactive tools in news article have become more and more sophisticated. It would be interesting to hear how you experiment with different approaches and would you think facilitate your readers or users the best. Also, how are these interfaces maintained and displayed to the reader? This topic will also include a discussion around how much should be facilitated towards the reader and how trust is maintained when presenting results from datasets.

Interview Questions – Phase 3

Mention that this field is unfamiliar to me and that I may use the wrong terms.

What is your role in this organisation?

How long have you been working for this organisation?

<Acquiring>

From where do you acquire data?

What sources do you use to acquire data? How official are these sources (crowdsourcing)?

What types of data do you acquire?

Is this data clearly defined by the data provider? If not, how do you handle this?

Does the data come with disclaimers, limitations or assumptions? If not, how do you ensure that you can trust this data?

How do you make sure this data is correct? Do you document this in any way?

<Documentation/Contextualisation>

How do you process this data? Do you compile it in a specific way?

Do you clean the data?

How do you document/contextualise this process?

What documentation is provided with the data?

<Accessibility>

Who reads your articles?

How do you present this data to your readers? Has this changed over the years?

Do you provide the original data? And if so, how do you store this and for how long?

How do you ensure trust in your organisation?

How do you engage with your readers?

Do you encourage your readers to re-use the provided data? How do you enable them to do this?

Is there an assumption made about the technical skills of the user?

What is your responsibility in explaining visualisation to the users?

Appendix 2 – Research methods

Transcription - Example Pages

Below a number of example pages of a transcript can be found. The highlighted links were additions added after contacting the participant about checking the transcription for any amendments.

Interview 004

Participant: Ray Moore, Archaeology Data Service

Location: King's Manor, University of York

File Name: Interview_Ray_Moore_Recorder.mp3

Date: 02/08/2019

Talboom: What is your role in the ADS?

Moore: Archives Manager, but also half of my time is spent as a digital archivist as well. So I do a bit of management but also day-to-day activity as well.

Talboom: What does Archives Manager entail?

Moore: It's looking after collections really and taking an overview of sort of the policy and procedure. Ensuring that the things that we do fit within standards of the profession and ensuring that we do things in the right way.

Talboom: How long have you been working for the ADS?

Moore: It's too long, too long.

[Laughter]

Moore: About ten years, eleven years. 2008 I started.

Talboom: To start of with the ingest of the data, what data do you collect and where does it come from?

Moore: We collect a broad gamut of research data, primarily from the archaeological and heritage environment. We preserve everything from reports and documents, right the way through to quite complex datasets; geophysics, 3-D data, databases, spreadsheets, images, drawings, like CAD drawings, photogrammetry, LiDAR, so there is a huge [scope] of different types of data that we collect. Essentially anything from the heritage, archaeological sector.

Talboom: And how does the data come here?

Moore: We have a variety of different submission streams. We have the ADS-easy submission stream, which is a semi-automated digital exchange of digital data. We also have a more informal exchange of digital data, we make use of external systems such as Dropbox and the University of York has a DropOff Service, which

is broadly similar [to Dropbox], so that people can essentially leave data for us. But also, sometimes the exchange of physical media as well, for particularly large datasets we still receive data on hard drives or USBs or pen drives.

Talboom: Is there any mandatory reason why depositors have to go through the ADS?

Moore: In some circumstances it is. Within the archaeological sector there is an increasing pressure from [within] the profession but also externally to try to be better about preserving digital data. Increasingly local authorities, for example, will make a requirement for people to send us digital data and we have formal agreements with some of those bodies in different locations throughout the UK [https://archaeologydataservice.ac.uk/research/partnerships.xhtml]. But also, within the research environment as well, often people as part of their funding applications are mandated to deposit data with us. So, we might receive the output of their research [as well]. We get data from both a research environment or people conducting research, but also within the commercial archaeological sector where people are conducting fieldwork and evaluations as part of their [daily] activities in advance of planning and the like.

Talboom: Are there any requirements from the depositors when they deposit data? So, formats, metadata, that kind of stuff?

Moore: Yeah, we have quite stringent policies and procedures in place to try and mitigate the problems that might arise in terms of the formats. So, we have particular formats that we expect, we have a list of accepted formats. They come in two forms, we have what we call our accepted formats, where there formats that we can take and our preferred formats, formats that we suggest people to deposit data in, because we have clear pathways in terms of preserving them, which makes it a bit easier for us in the long-term. We do have those requirements and also metadata requirements as well. We are quite stringent, we have quite extensive metadata requirements for depositors, so when people are depositing data through something like our ADS-easy portal they can complete that metadata online in a series of online forms, [or outside of the system they can also send metadata to us on spreadsheets/templates that we supply]. The extensiveness of our metadata is a problem and people do come back to us and say: 'These are quite extensive metadata requirements', [as we ask for] both collection level, in terms of documenting the dataset as a whole, and also in terms of documenting individual files and individual data objects as well.

Talboom: What happens if it is not in a required format? Or if the required metadata is not attached?

Moore: This is always a difficult one, we are quite strict in that sense. If the dataset arrives in a format that we don't accept, we will try and negotiate with the depositor about trying to get it in a format that we can accept really. Sometimes that's not always possible.

Talboom: Why would that be?

Moore: Say a depositor has died or receiving datasets from a project that has finished, or if a company has ceased trading, there might be issues there in terms of getting the data in formats that we want, but when that does happen we do continue to try and preserve that data, but it would be quite often on what we call a 'best efforts basis'. We can't really guarantee in the same way as we could do if we received the data in one of our preferred or accepted formats, but we can at least make the best effort that we can to preserve that in the future generations.

Talboom: When the depositor is selecting what they want to archive, do you have any role in that selection process?

Moore: Yes and no really. Wwe do provide guidance through our [Guidelines for Depositors] [https://archaeologydataservice.ac.uk/advice/selectionGuidance.xhtml] to try and assist people. Because quite often, in our experience, people are somewhat unsure about what they should include in their digital archive. It is so easy to create digital data, [but] people are [often] unsure about what the repository wants. So, for example, they might send us three or four versions of the same data that [they] have created at various stages in the project. Where possible we would like to receive the final version rather than having them deposit four or five versions of the same thing. It just fosters confusion, in our experience, [for] people who want to use data. Those four or five different versions are [confusing]. So we try and reduce that by providing guidance. But there's also stipulation within local authorities, for example, within the commercial sector, who would provide their own guidance in what they think the digital archive should contain. But we also work with external bodies like CIfA, which is the [Chartered] Institute for Archaeologists, they provide guidance to the profession, which is an attempt to make it more rigorous and more joined up. And one of their projects that we are involved in is their Selection Toolkit [https://www.archaeologists.net/news/archive-selection-toolkit-toolkit-aid-

selection-working-project-archive-1553864350], which is something that helps data creators to assess their own datasets and decide what the important things are. We also provide [specific] advice service ourselves, so if a depositor gets in contact with us and says: 'I've got this dataset…', we will provide help and assistance with that on a more personal basis. They might say: 'I've got this dataset and it consists of X, Y and Z, what's the best approach to preserving this?' [Or, more often, less specifically and more simply] 'what data do you want?'. So, we can [also] provide help and assistance with that.

Talboom: Stepping on from the ingest to the preservation. When the data comes to you, how is it then processed?

Moore: When it's received, we start by accessioning [the data]. We have an accession process by which we formally accept the data, so part of that is appraisal of the data. It's possible that at some point in the past that the depositor has shown us the dataset for us to [assess] and advise on what [is] the best [course of action], on what we want, or what might be the best things to preserve within a dataset. That process continues through the accession process. We have a look at the data, we make sure everything is okay, that the files aren't corrupted, that there aren't issues with the files and then we begin the process of creating metadata, technical metadata for those files. So we have a variety of different systems that we use, for example, the DROID software [http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/] which is an attempt to characterise data. So we will run that at accession so we can begin the process of creating checksums so we can ensure that that data remains unchanged throughout the history of it [life] at the repository. So, there's that, and we also check the metadata as well, as I said earlier, we have quite strict guidelines for our metadata, we look through that and assess it and make sure it is completed to the standards that we ask for. There might be a process of negotiation involved, so for example if metadata is missing, or if, for example, the file has been corrupted, we might have to talk to a depositor about those issues. But we try and do that as soon as we can after the data is sent to us. We try and accession the data within a short timeframe so it doesn't get forgotten, because quite often, in our experience, people send us data and if we don't respond in good time, people have moved on to other things or they have left organisations or given up the will to live after their research project is finished.

[Laughter]

We try and [do] that as quickly as we can after that. That accession process continues until we have received a signed deposit licence. When the data is deposited with us, we ask them to sign a deposit licence, which essentially gives us permission to disseminate and hold the data on their behalf and also to preserve that data. Once that happens, we then move through to a process of preserving the data, that involves the migration of data, the normalisation of data, into formats that we accept, that we use for preservation purposes.

Talboom: How do you select what formats they should be migrated to?

Moore: We have our own [data] procedures in place where we have assessed and reviewed over the twenty odd years that the ADS has been in existence, we asses the formats that we believe to be stable. We don't do that just by ourselves, we also engage with [our] community, the archaeological community, so we have our Guidelines for Depositors [https://archaeologydataservice.ac.uk/advice/guidelinesForDepositors.xhtml], which are based on…based on the sort of…what are they called? Essentially some guidelines that we have produced, I will remember the name as I go. We've engaged with our archaeological community to create these guidelines to sort of…what are they called?

Talboom: What ones are you talking about? The guidelines for depositors?

Moore: Yeah, the ones that Kieron works on.

Talboom: Is that the one with tDAR?

Moore: Yeah, that's exactly it.

Talboom: I can't remember either.

[Laughter]

Moore: I feel I should…

Talboom: Guide's to Good Practice?

Moore: The Guides to Good Practice [http://guides.archaeologydataservice.ac.uk/g2gpwiki/], that's it yeah. So, we have our Guides to Good Practice which are an attempt to garner the support of the wider archaeological community.

Talboom: But that's only for the archaeological community?

Moore: Yeah, that's specifically focused on the sort of data and datasets that archaeologists develop. But we also act as a broker in terms of considering the wider

archive…[and] digital preservation community, so we engage with that as well. Organisation[s] like the DPC who produce guidance [and watch reports] on [formats and data types] and the like. Also what other [archives] are doing, like the National Archives, to see [how] they are preserving data. These feed in to our own internal procedures and our own guidelines and they really allow us to create those [lists of] stable formats that we think will have longevity for future preservation of data.

Talboom: Then you have migrated the file, what happens then?

Moore: Once we have normalised data, we then move forward to creating the AIP, which is the Archival Information Package. Which is essentially the preservation version of the dataset, but we also create the DIP, which is the dissemination information package, which is the stuff that we disseminate for our archive.

Talboom: And is that then in a different format?

Memo – Example

This memo was named *2020_11_24_Thoughts_After_Coding_Second_Round.docx.* All memos were named in a similar manner to make it easy to find a memo and know when it was written.

Thought after coding

It seems to be splitting into three different area, and then there is also an area with a bunch of codes that don't really fit anywhere.

But it seems to be:

Role of the archivist – This is around what we should and should not be doing and how there should be a stronger divide in what the archivists does and the developer, which is very unclear at the moment and completely understandable due to the nature of the material. Also, context has to be different for this material, this is a strong point of archivists. Another thing is the ethical side of everything, which archivists should spend more time thinking about: It's not technical, it's ethical.

What is our responsibility? – This is about where we draw the line with our users. This also comes into who our users are and how this is difficult in this setting, as explained in earlier points, it's a real struggle. But also where it is that we should draw the line, and this is where the developers are very clear in it. They just have a very different approach to it, which we should learn to partially adapt.

Broader changes in the sector – This is not necessarily tied to any one of these and is more around our broader work in the archives. The idea of working together is important here, which is highlighted by a number of participants. But also around helping smaller archives who may not be fortunate enough to provide this type of network. Building tools for everyone, not bespoke tools and also funding this differently. But also being more like Big Tech and using what is out there, open-source tools are great for certain infrastructures.