

University of Southampton Research Repository

Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.

When referring to this thesis and any accompanying data, full bibliographic details must be given, e.g.

Thesis: Author (Year of Submission) "Full thesis title", University of Southampton, name of the University Faculty or School or Department, PhD Thesis, pagination.

Data: Author (Year) Title. URI [dataset]

UNIVERSITY OF SOUTHAMPTON

DOCTORAL THESIS

Saving the Web: Facets of Web Archiving in Everyday Practice

by

Jessica Ogden

*A thesis submitted in fulfilment towards
the degree of Doctor of Philosophy*

in the

Faculty of Social Sciences
Sociology, Social Policy and Criminology

ORCID iD: <https://orcid.org/0000-0003-4696-7340>

August 2020

UNIVERSITY OF SOUTHAMPTON

Abstract

Faculty of Social Sciences

Sociology, Social Policy and Criminology

Doctor of Philosophy

Saving the Web: Facets of Web Archiving in Everyday Practice

by Jessica Ogden

This thesis makes visible the work of archiving the Web. It demonstrates the growing role of web archives (WAs) in the circulation of information and culture online, and emphasises the inherent connections between how the Web is archived, its future use and our understandings of WAs, archivists and the Web itself. As the first in-depth sociotechnical study of web archiving, this research offers a view into the ways that web archivists are shaping what and how the Web is saved for the future. Using a combination of ethnographic observation, interviews and documentary sources, the thesis investigates web archiving at three sites: the Internet Archive – the world’s largest web archive; Archive Team – ‘a loose collective of rogue archivists and programmers’ archiving the Web; and the Environmental Data & Governance Initiative (EDGI) – a community of academics, librarians and activists formed in the wake of the 2016 US Presidential Election to safe-guard environmental and climate data. Through the application of practice theory, thematic analysis and facet methodology, I frame my findings through three ‘facets of web archiving’: *infrastructure*, *culture* and *politics*. I show that the web archival activities of organisations, people and bots are both historically-situated and embedded in the contemporary politics of online communication and information sharing. WAs are reflected on as ‘places’ where the past, present and future of the Web collapses around an evolving assemblage of sociotechnical practices and actors dedicated to enabling different (and at times, conflicting) community-defined imaginaries for the Web. WAs are revealed to be contested sites where these politics are enabled and enacted over time. This thesis therefore contributes to research on the performance of power and politics on the Web, and raises new questions concerning how different communities negotiate the challenges of ephemerality and strive to build the ‘Web they want’.

Contents

Contents	v
List of Figures	ix
List of Tables	xi
Declaration of Authorship	xiii
Acknowledgements	xv
1 Introduction	1
1.1 Prologue	1
1.2 Web Archiving	4
1.2.1 The Ephemeral Web	4
1.2.2 Engaging the Archived Web	6
1.3 ‘Opening the Black Box’ of Web Archiving	9
1.4 Aims of the Thesis	12
1.5 Chapter Summaries	14
2 Archiving the Web, Mapping the Field	17
2.1 Web Archiving in Practice	17
2.1.1 Key Initiatives	18
2.1.2 Components and Technologies of Web Archiving	21
2.2 Problematising Web Archival Practices	27
2.2.1 Technological Challenges	28
2.2.2 Legal and Ethical	29
2.2.3 Defining, Selecting and Using the Object of Collection	31
2.3 Framing the Materiality of Web Archiving	33
2.3.1 The Digital Turn in Archives	35
2.3.2 Performative (Web) Archives	38
2.4 Chapter Summary	41
3 Observing Web Archiving: A Methodology	43
3.1 Methodological Design	43
3.1.1 The Qualitative Approach and Paradigm	44

3.1.2	A Theoretical Framework: Practice Theory	45
3.1.3	Ethnographic Methods in Practice	47
3.2	Selecting the Sites of Investigation	51
3.3	Data Collection	56
3.3.1	Ethnographic Interviews	56
3.3.2	Participant and Non-participant Observation	59
3.3.3	Documentary and Secondary Sources	62
3.3.4	Implications of the Approach	62
3.4	Thematic Analysis and the Facet Approach	69
3.5	Chapter Summary	72
4	Web Archiving as Infrastructure: The Internet Archive	75
4.1	Introduction	75
4.2	Locating Infrastructure, Framing the Internet Archive	77
4.2.1	Brewster Goes West and the Dot-com Boom	78
4.2.2	The ‘Spiritual Centre’ of Silicon Valley	81
4.3	Infrastructure as Labour	84
4.3.1	Knowledge Work	85
4.3.2	Translation	91
4.3.3	Breakdown, Maintenance and Repair	95
4.4	Chapter Summary	99
5	Web Archiving as Culture: Archive Team	101
5.1	Introduction	101
5.2	Framing Archive Team, Constructing Community	103
5.2.1	An ‘Emergency Response Team’ for Web Archiving	105
5.2.2	Archive Team as Community Protocols	107
5.3	Archive Team at Work	116
5.3.1	Yahoo and the Case of Tumblr	116
5.3.2	Distributed DIY Web Archiving	119
5.3.3	Transforming Culture	127
5.4	Chapter Summary	136
6	Web Archiving as Politics: Environmental Data & Governance Initiative	139
6.1	Introduction	139
6.2	Origins of EDGI, Framing a Crisis	141
6.2.1	The Threat of Donald Trump and the Harper Years	143
6.2.2	Uncertainty, Urgency and the End of Term	146
6.3	DataRescue and the Boundaries of Web Archiving	149
6.3.1	Mobilising Publics and Expertise	151
6.3.2	Collaborative Interventions	157
6.4	Reflections on the Politics and Impact of DataRescue	170
6.5	Chapter Summary	174

7	Conclusions and Future Work	177
7.1	Facets of Web Archiving: A Review	177
7.2	Facets of Web Archiving: Thesis Contribution	181
7.3	The Web They Want	187
A	Participant Information Sheet - Organisations	191
B	Participant Information Sheet - Online Community	195
C	Participant Information Sheet - EDGI	199
D	Consent Form - Organisations	203
E	Consent Form - Individuals	205
	Bibliography	207

List of Figures

2.1	Archive-It partner organisation types, as tallied from the Archive-it collection website	20
2.2	The <i>Web Archiving Life Cycle Model</i> for best practices in web archiving (Bragg and Hanna, 2013, p.3).	22
2.3	The Wayback Machine with keyword search	26
2.4	The Wayback Machine results page indicating the number of ‘snapshots’ that were taken on the day	27
3.1	Model of the levels of participation in observation, as adapted from Spradley (1980, p.58).	60
3.2	The process of rapport development from Spradley (1979, p.79), as exhibited by informants and participants in the ethnographic research process.	66
3.3	A cut gemstone and its facets.	71
4.1	Brewster Kahle and Bruce Gilliat in 1998 at the headquarters of Alexa Internet	79
4.2	The Head Office of the Internet Archive at 300 Funston Avenue, San Francisco, California	82
4.3	A selection of the 100+ ceramic statues of Internet Archive staff that have worked at the Archive for at least three years	83
4.4	Petabox servers in the rear of the Great Room at the Internet Archive .	84
4.5	Brewster Kahle, founder of the Internet Archive, speaking at their 20 th Anniversary Party	86
4.6	Internet Archive catalogue tasks, in process	93
4.7	The Internet Archive Weather Map shows network traffic in and out of each component of the server infrastructure	96
5.1	Archive Team, a rogue band of activist web archivists formed in response to the shuttering of online hosting services for user-generated content	105
5.2	A screenshot from the Archive Team wiki showing basic information associated with the Tumblr project	110

5.3	A screenshot from the Archive Team wiki showing a selection of Warrior Projects, organised by the most recent project	111
5.4	The rate of Archive Team collection activities between 2009 - 2018, in terabytes over time	120
5.5	The Warrior application downloading Tumblr NSFW blogs	122
5.6	Graphic depiction of the Archive Team's Warrior web archiving infrastructure	123
5.7	Screenshot from the Archive Team Tumblr NSFW Tracker, displaying the uploaders and uploads (still in process)	124
5.8	Screenshot of Jason Scott tweets used to mobilise participants in Archive Team's efforts to archive Tumblr NSFW posts	126
5.9	Two posts that Archive Team promoted on Tumblr to generate submissions and support for the web archiving project.	128
6.1	A map of DataRescue events between December 2016 - June 2017 . . .	150
6.2	A <i>Washington Post</i> article that included coverage of the 'guerrilla archiving' event	156
6.3	A working screenshot of the EDGI Google Chrome extension used to submit seeds to the 2016 End of Term project	161
6.4	A screenshot of an EPA Agency Primer used as part of the DataRescue workflow	163
6.5	Spectrum of data archiving at DataRescue (Allen et al., 2017)	165
6.6	Screenshot of home page of the Archivers.space application	166
6.7	Screenshot of an item in the Archivers.space application	167
6.8	A graph of the 2016 End of Term nominations over time	171
7.1	The Internet Archive's web exhibition of the 1996 US Presidential election web archive	185
7.2	Screenshot of Twitter user reaction to Archive Team web archiving Tumblr NSFW	186

List of Tables

4.1	Crawl events at the Internet Archive	87
5.1	Archive Team Warrior project categories of success	109

Declaration of Authorship

I, Jessica Ogden, declare that this thesis titled, '**Saving the Web: Facets of Web Archiving in Everyday Practice**' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University;
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- Where I have consulted the published work of others, this is always clearly attributed;
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- Parts of this work have been presented and/or published as:

Ogden, Jessica, Susan Halford, and Leslie Carr. 2018. 'Re-Configuring Web Imaginaries: Reflections on Community-Based Web Archiving for Data Justice'. Conference Presentation presented at the British Sociological Association Conference 2018, University of Newcastle, Newcastle, UK, April 10.

Ogden, Jessica, Susan Halford, and Leslie Carr. 2017. 'Maintaining the Web: Web Archiving, Labour and the Internet Archive'. Conference Presentation presented at the Association for Internet Researchers Conference 2017, Tartu, Estonia, October 20. Extended abstract published by *AoIR Selected Papers of Internet Research* (2019). <https://spir.aoir.org/ojs/index.php/spir/article/view/10186>

Ogden, Jessica, Susan Halford, and Leslie Carr. 2017. 'Observing Archives: Web Archiving as sociotechnical Practice'. Conference Presentation presented at the 4S: Social Studies of Science and Technology Conference 2017, Boston, MA, USA, September 2.

Ogden, Jessica, Susan Halford, and Leslie Carr. 2017. 'Observing Web Archives: The Case for an Ethnographic Study of Web Archiving'. In *Proceedings of WebSci'17, Troy, NY, USA., June 25-28, 2017*. ACM. <https://doi.org/10.1145/3091478.3091506>.

Ogden, Jessica. 2016. 'Interrogating the Politics and Performativity of Web Archives'.
Doctoral Consortium Presentation presented at the Joint Conference on Digital Lib-
raries 2016: Doctoral Consortium, Rutgers University, New Jersey, June 19.

Signed:

Date:

Acknowledgements

This research has benefitted from the support and contributions of many people over the course of the (long) journey that deserve recognition.

This work was funded by the UK Engineering and Physical Sciences Research Council and the Web Science Centre for Doctoral Training, Grant No. EP/G036926/1. To my supervisory team, Prof. Susan Halford and Prof. Leslie Carr, thank you. In particular, thanks to Susan whose support and critical engagement has been unwavering throughout. Thanks for getting me over the finish line.

Thanks to the many participants in this research, including the staff and volunteers at the Internet Archive, Archive Team and the Environmental Data & Governance Initiative. Thanks for opening your doors (physical and virtual) and for allowing me to hang around – this work would not have been possible without all of you. Thanks to the WASAPI project for funding my travel to the Archive.

Thanks to my many Web Science DTC/CDT colleagues. *WebSciC5 4-eva*. In particular, to Faranak Hardcastle, Jack Webster and the three bears, Conrad D’Souza, Joanna Munson and Michael Day.

Thanks to the Archives Unleashed team for letting me come out and spitball web archives for three years. I can’t imagine having done this research without the opportunity to meet and work with other practitioners and grad students working in this space, and for that I am so grateful. In particular, thanks to my partners in crime, Mat Kelly, Shawn Walker and Ed Summers. To Shawn and Ed, our SPN Friday afternoon chats have been invaluable. Thank you for all that you have no doubt contributed to this work. Thanks for listening and joining me on my web archiving flights of fancy.

To my L - P colleagues who reluctantly let me abscond away to do this Web Science thing, I am grateful.

To Tenz, thanks for the letters, the never-ending playlist, a couch to crash on and the will to keep at the writing. To my hometown girls Katie, Amanda, Ashley, Lesley and Liz – love and thanks for the support from afar.

Many thanks go to my furry sidekick, Rosie. This journey has been immeasurably better thanks to the many many walks, cuddles and games of tug-o-war you demanded along the way.

And finally, thanks to my family near and far who have been my rock through what have most certainly been the toughest years yet. Love and gratitude to my siblings Nicole, James and Jon for the yoga stretches, the pep talks and general distractions. To my Mom, everything I am, I owe to you.

Particular love and gratitude goes to my partner in life, Leif Isaksen. You've been right by my side during all of the many ups and downs presented in the last five years and I can't possibly imagine making it through without your love and support. Also, you make the best cups of tea.

THANK YOU.

To my father. With love, always and forever.

1

Introduction

1.1 Prologue

In April 2018, Twitter and web archivists alike erupted in reaction to the news that Joy-Ann Reid, a morning talk show host and journalist at MSNBC (a US-based broadcast news company), had accused the Internet Archive of archiving ‘altered’ blogposts from her political pundit blog *The Reid Report*. Reid stood accused of authoring a series of incendiary posts in the early 2000s that attempted to ‘out’ supposedly closeted homosexual politicians, supported homophobic and anti-gay marriage views, as well as implicated others such as retired NBA player Tim Hardaway as a homophobe. Philip Bump, a journalist writing for the *Washington Post*, explains how the Twitter user @Jamie_Maz made headlines:

“@Jamie_Maz found the article because it had been indexed by the Internet Archive’s Wayback Machine, a mostly automated tool that, for more than two decades, has been storing copies of Web pages as they appeared at the time. At some point in 2007, the Internet Archive’s system visited Reid’s blog and saved a copy of the Hardaway article, dropping it in a database where it sat mostly ignored for 11 years — until @Jamie_Maz uncovered it” (Bump, 2018).

The news cycle that dominated thereafter included various interpretations of how these blogposts came to be archived at the Internet Archive, including a statement from Reid that after consulting cyber-security experts they had determined that ‘an unknown, external party accessed and manipulated material from [their] now-defunct

blog’ (Ecarma, 2018). Responding to the allegations of hacking, the Internet Archive issued a response on their blog, an excerpt of which reads:

“When we reviewed the archives, we found nothing to indicate tampering or hacking of the Wayback Machine versions. At least some of the examples of allegedly fraudulent posts provided to us had been archived at different dates and by different entities.

We let Reid’s lawyers know that the information provided was not sufficient for us to verify claims of manipulation. Consequently, and due to Reid’s being a journalist (a very high-profile one, at that) and the journalistic nature of the blog archives, we declined to take down the archives” (Butler, 2018).

The use of the archived posts from *The Reid Report* provides a window into some of the ways in which citizens and journalists are leveraging the archived Web to support their claims and hold political and public figures to account. Further, this example concisely demonstrates several aspects of web archiving that are of central concern to this thesis. First, the case of *The Reid Report* clearly illustrates the functional role that web archives – and the Internet Archive Wayback Machine (IAWM), in particular – are playing in the performance of web-based *informational politics* (Rogers, 2004), through their provision of historical records for deleted and otherwise dynamic web content. Here, the web archive is valued as a source for historical information through time; exemplified through @Jamie_Maz’s use of the IAWM to access numerous blog-posts that have long been deleted from the ‘live Web’. The implied importance of the IAWM is also clearly expressed through the *Washington Post* headline: ‘The Joy Reid fight reinforces how critical the Internet Archive is to modern politics’.

Second, this example highlights popularised conceptions about the role of automation in web archiving, whilst *black boxing* unknown and invisible processes that enable web archiving in practice. The language used in the above excerpt from the *Washington Post* article describing the backstory behind @Jamie_Maz’s discovery is instructive. Here, the IAWM is rendered as an unbiased agent and neutral receptacle; as a dumb ‘mostly automated tool’ that stumbled onto Joy-Ann Reid’s blog and fortuitously made a copy. It portrays the IAWM as an unknowing place where articles are ‘saved’ then ‘ignored’, but also one where they can be ‘found’ and ‘uncovered’ by willing users. This passive language of automation is reminiscent of others who have likened the IAWM to a ‘lobster trap’ that ‘[sits] passively in the ocean, placed in areas of strategic interest’ (Karpf, 2012). However, as Ben-David and Amram (2018) have similarly problematised the ‘epistemic assumptions’ embedded in the depiction of the IAWM as a passive agent, this portrayal invariably obscures the complex assemblage of sociotechnical relations at work in web archives. This is further supported through a wide range of relevant scholarship that challenges notions of ‘naturally

occurring' archives and makes the case for acknowledging (and studying) the implications of invisible labour and the human component of information infrastructures (Star and Ruhleder, 1996), data production (Ribes and Jackson, 2013) and memory work (Arnold, 2016).

Setting aside the hacking allegations, the excerpt from the Internet Archive's response to Reid alludes to some of the everyday decisions that are made by web archives about what is archived and how these archives are made available. The statement complicates the 'lobster trap' metaphor and gestures to multiple mechanisms by which web pages end up in the IAWM. They indicate that 'at least some of the [posts]' were provided at 'different dates and by different entities', however, at the time there was no way to know who and what those entities were due to a temporary policy decision to remove public access to the *The Reid Report* snapshots. This highlights problems with how users of the IAWM (and web archives, more generally) are to contextualise and understand the origins of archival collections – *who* archived a website, *how* and *why*?

In addition to the unknown dimensions of collection practices, the Internet Archive's statement raises additional questions about how decisions are made surrounding what archives are made accessible and when. The IA's assertion that Reid's position as a 'very high-profile' public figure alongside the 'journalistic nature' of their blog points towards an unknown set of criteria for determining public access. Therefore as a final point, this example also underscores the fundamental operation of power when web content moves between the live Web and web archives. Here, as content moves it becomes subject to new *regimes of control*, interfaces and access protocols that dictate how the potential use of web archives is enabled and constrained.

In summary, this example reveals both the importance of web archives in the face of a dynamic Web and why web archives deserve further exploration. As shown in the Joy-Ann Reid example, web archives such as the Internet Archive provide a growing resource for accessing lost web content. These resources are subject to diverse and largely unknown collection techniques that are often portrayed as automated, unknowing, neutral and objective. Like other complex sociotechnical systems that involve both human and semi-automated agents on the Web, web archives and the practices that underpin how they function are not widely understood – particularly in contexts that operate outside of conventional memory institutions historically tasked with maintaining historical records. In this thesis I focus on just this: *how does web archiving actually get done?*

As the Web itself is increasingly becoming 'the archive' of everyday life, web archives offer an opportunity to capture and temporarily stabilise a view of life online at an unprecedented scale (Milligan, 2019). Given this scale however, any attempts to capture the Web will always be partial. The ways in which web archives are collected and maintained, therefore, have important implications for what of the Web is saved

and indeed its future knowability. By expanding and deepening knowledge of web archiving in practice, this thesis therefore interrogates the operation of power and politics surrounding how and what of the Web ‘gets saved’, for whom and for what purpose.

1.2 Web Archiving

The remainder of this chapter works to further frame the motivations for this thesis. First, I take a step back to consider the ways that the ‘ephemeral Web’ has led to the initial creation of web archiving initiatives. Second, in an effort to underscore the growing role of web archives in the circulation of information and culture online, I detail some of the ways that web archives are being used by scholars, citizen activists and others to mitigate the effects of the dynamic Web. Here, I outline three fundamental concerns associated with web archives in order to emphasise the inherent connections between *how* the Web is archived, its future use and our understandings of web archives, archivists, and by extension, the Web. Next, I make the initial case for a sociotechnical approach to studying web archiving through a brief discussion of existing research into web archiving practices (that will be extended in Chapter 2). This chapter ends with a discussion of the aims and approach of this thesis and finally, a brief outline and summary of the remaining chapters.

1.2.1 The Ephemeral Web

The Web’s transience, at least in part, can be attributed to the underlying technical affordances of the web architecture which inherently relies on a decentralised and distributed hypertext system. The persistence and resolvability of hyperlinks based on *universal resource locators* (URLs) are not guaranteed, therefore leading to the now commonplace ‘404 Not Found’ experienced when web users attempt to access URLs that have either moved or are no longer live. Concerns about the ephemerality of URLs have been reinforced by various longitudinal studies of ‘link rot’ or ‘link decay’ (Oguz and Koehler, 2015), where the average half-life of web pages and websites has been estimated anywhere between 75 days (Kahle, 1997) and two years (Koehler, 2004). This is bolstered by social media research (SalahEldeen and Nelson, 2012) which found that during a period between 2009 and 2012, on average, 11% of online resources shared on social media failed to resolve one year later.

And so, since the mid-1990s, institutions such as national libraries and the Internet Archive have been ‘archiving the Web’ by harvesting and preserving the live Web in web archives. Web archiving has roots in a wider digital preservation movement which emerged in the 1980s-1990s, led by memory institutions to develop strategies for addressing the rise of personal computing and the impact of digital artefacts on their abilities to capture and preserve ‘records of social phenomena’ (Schneider and

Foot, 2008). This was particularly fuelled by fears over the so-called ‘digital dark ages’, a term first used by Kuny (1997) to describe a scenario where the development pace of technologies (used to produce digital objects) outweighs that of the investment in infrastructures, technologies and policies to preserve them long-term. As the world’s information and communication platform is increasingly born-digital and online, web archives have been positioned as key to capturing and preserving contemporary digital cultural heritage, to ensuring stability and access to pre-existing web resources and facilitating new knowledge via scholarly research. Web archives in their various forms – including social media archives – have become a sort of ‘prosthesis’ for the Web and a necessary pre-condition for any research into the Web(s) of the past and near-present.¹

In practice, web archiving initiatives have ranged from the large-scale activities of national libraries and archives (like the British Library and UK National Archives), the Internet Archive and the work of networked communities such as Archive Team;² to the individual efforts of scholars creating web archives for their own purposes. Surveys of institutional practitioners reveal a growing number of organisations that are web archiving, albeit with limited staff-time resources and with the majority of respondents using some form of external service to collect or manage their web archives (Bailey et al., 2014, 2017; National Digital Stewardship Alliance, 2012). Much of the focus of the field has been on the continued development of technologies and standards for web collection development (Hockx-Yu, 2014a), with increased attention on facilitating the scholarly use of web archives (Dougherty and Meyer, 2014; Dougherty et al., 2010; Meyer, Thomas and Schroeder, 2011).

As such, a number of projects and publications in recent years have been instrumental in outlining both the opportunities and the challenges of engaging with web archives from the perspective of scholars and practitioners.³ Further to this, recent geo-political events coupled with a growing public awareness about web archives has given rise to a number of use-cases, community groups and projects that stretch beyond just academic engagement with web archiving and the archived Web. These encounters (both scholarly and elsewhere) have underscored the role of web archives in the circulation of information and culture online, but have also worked to implicate web archiving itself in shaping claims about the past Web. Below, I further elaborate on these points in order to outline the inherent connections between the ways the Web is archived and its possibilities for capturing the Web for the future.

¹Inspiration for this analogy is taken from Derrida’s (1998) treatment of ‘technological devices for archiving’ as prostheses for memory formation and storage.

²<http://www.archive-team.org> (visited on 9th Jul. 2019)

³For a recent discussion on the state of the field, see Taylor’s (2017) Introduction to the *Journal of Western Archives Web Archiving* Special Issue.

1.2.2 Engaging the Archived Web

Web archives for scholarly use

The opportunities presented by web archives as a tool for ‘web historiography’ have been widely relayed by historians and communication scholars (Brügger, 2012, 2013; Brügger and Milligan, 2019; Foot and Schneider, 2010; Milligan, 2019). Early work in the field by Schneider and Foot (2004) describes various methodological strategies for using web archives for the benefit of academic research. Here Schneider and Foot outline methods that focus on the Web itself as an object of study, as well as ‘sociocultural’ approaches that take into account the ‘hyperlinked contexts and situatedness of websites’ in what they call ‘web sphere analysis’ (Foot and Schneider, 2006). Examples of studying the historical growth of the Web highlight the opportunities presented by domain-scale analyses of web archives, including research into: the growth and evolution of the Danish domain names (Brügger, Laursen and Nielson, 2017), the evolution of the UK web domain and UK universities on the Web (Hale et al., 2014; Meyer et al., 2017), the death of the Yugoslavian web domain (Ben-David, 2016) and the numerous studies of national web domains outlined in a recent edited volume by Brügger and Laursen (2019). ‘Web spheres’ analysis is exemplified through, for example, the study of early Web platforms such as *GeoCities* (Milligan, 2017) – where the (now defunct) platform and the communities within them demarcate the boundary of study.

By extension, web archives therefore also present opportunities for examining the recurring themes in Internet research (and the social sciences, more generally) by providing temporarily stable resources for the study of issues related to online ‘identity, community, inequality, politics, organizations, and culture’ (DiMaggio et al., 2001; Sandvig and Hargittai, 2015, p.6). Here research has often been driven by more thematic interests, where the construction of targeted collections enables the study of, for example, the everyday life of French migrant communities in London (Huc-Hepher, 2015), local news production in the United States (Weber and Napoli, 2018), or the historiographical study of web design during the dot-com era (Ankersson, 2015, 2018), to name only a few.

Web archives for evidence-based accountability

In addition to the increased promotion and use of web archives as a source for scholarly research, there is evidence that a diverse range of stakeholders are engaging with the archived Web, with a growing number of examples, in particular pointing towards the use of web archives as tools for evidence-based accountability. For example, since 2003, lawyers and legal teams have regularly used the IAWM to evidence intellectual property claims (Eltgroth, 2009) – a practice that was further reinforced by another

recent ruling by a US Court of Appeals that made Wayback snapshots admissible as evidence.⁴

As illustrated in the Joy-Ann Reid example (discussed in Section 1.1), web archives provide access to websites and social media that may be censored, restricted or deleted from the live Web, as well as providing sources for journalists, civic technologists and citizen activists to hold public figures to account. For example, *Politwoops* provides rolling public access to the deleted tweets of politicians (as well as statistics on tweet and delete patterns),⁵ and during the 2013 US Government shutdown, the IAWM became the only accessible source for core government websites and information services (Kahle, 2013). Other examples include the ways that the ‘snapshot’ approach – or the systematic capture of the same URLs over time – can be used to demonstrate ‘content drift’ (Klein et al., 2014) or changes to web-based content. For example, in the wake of the US Presidential election of Donald Trump in 2016, web archives have been regularly used by activists, journalists and the mainstream media to fact check and detect changes to: official documents and transcripts released by the Whitehouse (Watson, 2018) federal environmental policy agendas associated with climate change (Wallace, 2018), the availability of immigration and asylum officer training documents (Campbell, 2018, 2019) and information related to healthcare and the 2010 Affordable Care Act (Bergman et al., 2019) – to name only a few.

Two further examples are frequently cited when making the case for the importance of collecting web archives as records of the past Web. The first example stems from 2013 when the UK Conservative Party deleted ‘a decade’s worth’ of political speeches previously hosted on their website (Ramesh and Hern, 2013). The deletion was accompanied by the use of the `robots.txt` protocol to block crawlers and access to archived versions of the site contained on the IAWM – a move that for some, signalled the political motivations underlying the deletion.⁶ As such the Tory party’s deletion was met with extensive media coverage, including from *Computer Weekly* who broke the story (Ballard, 2013) and emphasised ‘how fragile the historic record is on the Internet’ (Ramesh and Hern, 2013).⁷ In addition, however, the story also worked to publicly draw attention to the existence of targeted collections by the UK Web Archive (UKWA), who have been archiving the party website (with their permission)

⁴Since 2003, IAWM snapshots have been admissible as evidence in the US court of law (Eltgroth, 2009), often requiring Internet Archive representatives to provide expert testimony regarding web archiving procedures and the validity of the archives. Further precedent was made by a recent ruling by a US Court of Appeals Second Circuit in 2018 to accept them as evidence (McCarthy, 2018).

⁵<https://politwoops.eu/> (visited on 9th Jul. 2019)

⁶The `robots.txt` protocol is a web standard and set of guidelines used to communicate between web masters and web crawlers (Koster, 1993). Web masters place a `robots.txt` file in the root of their host directory which can contain a series of instructions that allow or disallow certain named web crawlers from crawling their website or particular parts of a website. The IAWM also uses the `robots.txt` to determine access rights, and in the case of the Conservative Party website, the interpretation was that they deliberately changed the `robots.txt` to disallow public access to the archived website in the Wayback Machine.

⁷Ballard (2013) likened the Tory party’s deletion to ‘men in black’ taking history books from a public library and burning them in the car park.

since 2004 (Webster, 2013). Therefore, in lieu of use of the robots.txt protocol to block public access to the IAWM, the UKWA was for a time, the only way to access ten years worth of Conservative Party speeches.

The second example occurred in 2014, when the IAWM was used to support allegations surrounding the downing of Malaysia Airlines Flight 17 in Ukraine (Dewey, 2014). Here, the IAWM's public 'Save Page Now' feature⁸ was used to archive a social media post by a pro-Russian Ukrainian separatist claiming they had shot down a plane (which turned out to be flight MH17).⁹ The case of MH17 provides a 'powerful testament' (Dewey, 2014) for the role of web archives in the circulation of information online, where the archived post contributed to the investigation and eventual condemnation of Russian forces by the international community.

Framing the significance of web archiving practices

The broad use of web archives by both scholars and the wider public (e.g. journalists, citizen activists and lawyers) has positioned both web archiving and web archives as necessary and legitimate sources in the face of an ephemeral Web. However, with this increased engagement, three fundamental points of concern are raised that are central to the framing of this thesis.

First, web archives are contingent on practice. This may appear to be an obvious point to make, however, it draws attention to the notion that rather than an inevitable outcome of the Web's ephemerality, use and access to the archived Web is contingent on the interventions of human and technical actors to 'save the Web'. This point has been made elsewhere by historians and communication scholars who argue that the Web as a information/communication medium has fundamentally impacted the nature of the historical record and scholars' ability to study the past (Brügger, 2018; Milligan, 2019). Joy-Ann Reid's blog, the Conservative Party website and the Ukrainian separatist tweet all had to be archived in order for researchers, journalists and citizen activists to be able to make claims about what was said and done online. The purpose here is not to assert value judgements about whether or not these web resources should have been archived, but rather it is to argue that these preservation interventions have enabled a set of claims to be made about the World that would otherwise be impossible given the medium through which they were originally communicated.

Second, web archives are inherently shaped through practice. Here, as Brügger (2018) has argued, web archives should be seen as ontologically different to both

⁸Save Page Now is a set of tools hosted by the Internet Archive that enable public users to archive and host web pages in the Wayback Machine. Further details about Save Page Now can be found in an Internet Archive blog promoting the use of the tool at: <https://blog.archive.org/2017/01/25/see-something-save-something/> (visited on 10th Jul. 2019).

⁹https://web.archive.org/web/20140717155720/https://vk.com/wall-57424472_7256 (visited on 10th Jul. 2019; archived on 17th Jul. 2014)

the Web and conventional conceptualisations of 'archives' (analogue, digital or born-digital). Distinguished from other 'born digital' material, Brügger (2016) characterises web archives as 're-born digital' objects that are fundamentally shaped by the processes undertaken during their collection, preservation and access. Here a combination of both collection decisions and 'technical problems' leads to web archives that are not *copies* of the live web, but rather *contingent constructions* where 'the process of archiving itself may change what is archived, thus creating something that is not necessarily identical to what was once online' (Brügger, 2012, p.108). Foregrounding the subjective nature of practices, Brügger (2009, 2013) outlines how choices surrounding what to archive, what strategies and software are employed, how archives are re-assembled for use, as well as problems that may occur during archiving – all create unique representations of the Web. Furthermore, these observations also underscore the fact that any attempts to archive the Web will always be partial and incomplete. Web archives are therefore, particular representations of the Web – they are inherently situated and dependent upon the mechanisms of their production.

Third, unpicking practice is key to understanding web archives, archivists and by extension, the Web. In the first instance, the point here is that the 're-born' status of web archives creates an imperative for users of web archives to employ what has been referred to as 'digital source criticism' (Rogers, 2017), or the practice of critically engaging with the nature and validity of the source material used to make claims about the world. Although practices associated with source criticism have been further complicated by the sociotechnical complexity of web archives (Ben-David and Huurdeman, 2014; Brügger, 2012; Rogers, 2017), by unpicking the practices undertaken to create and re-construct the archived Web, users better apprehend their affordances and the impact practices have on the nature and possible limitations of future claims made using web archives. As an extension to this point however, I want to argue that engaging with web archiving practices also tells us something about web archival practitioners and the Web(s) they seek to archive. By engaging with the myriad of motivations, value-statements, mechanisms and circumstances under which web archivists operationalise their aims, we learn about how the Web itself is valued, situated and framed within the everyday lives of practitioners. Studying practices therefore, critically re-centres and acknowledges both the humans and technical agents embedded within the processes of saving and engaging with the Web's past.

1.3 'Opening the Black Box' of Web Archiving

Rather than frame web archiving as an inevitable outcome of an ephemeral Web, this thesis explores the ways that web archiving is rooted in particular cultural worlds and practices that have implications for how and what of the Web is archived. As such, this thesis draws on (amongst others) a social constructivist view of web archiving 'to

emphasise the contingency and choice rather than forces of necessity' (Winner, 1993, pp.366-367) involved in the enactment of web archiving. Here, the aspiration of 'opening the black box' is not necessarily to 'know all' but rather, as others have suggested (Bucher, 2012; Chun, 2013), to make sense of the diverse material relations that make up sociotechnical systems and the knowledge they produce.

Other studies have attempted to engage with web archiving practices from different theoretical and methodological perspectives, some of which I briefly outline below in order to situate the contribution of this thesis within the landscape of previous research. Overall, previous work can be divided into overlapping clusters of research that draws on different methodological approaches and aims for studying web archiving practices. These clusters of research offer overviews of the field and contemporary practices through time, studies that attempt to address the gap between the collection of web archives and their use by scholars, and research that seeks to 'reverse engineer' or reconstruct practices through 'forensic' studies of web archives themselves. Although each of these clusters complements the approach taken here, this thesis aims to both widen and deepen an understanding of web archiving through the use of ethnographic methods and by focusing on the socio-materiality of web archiving practices across different sites (discussed in the next section).

Various windows into practice have been published that attempt to map the field of web archiving, including overviews of the field (Brown, 2006; Masanès, 2006), literature reviews (Niu, 2012) and 'best practice' documents (Bragg and Hanna, 2013; Day Thomson, 2016; Pennock, 2013) that provide models for charting the general components of web archiving. However, some of this documentation has failed to keep up to date with the pace of web/archiving technology development, and is often limited to specific software, tools and professional contexts (e.g. library/bibliographic approaches to web archiving). Additional research has sought to get a view of the types of organisational contexts in which web archiving is taking place. Here, efforts have been made to establish a baseline for knowledge pertaining to who is web archiving, the tools and technologies they are using and to a lesser extent, what they are archiving (Bailey et al., 2014, 2017; Truman, 2016; Vlassenroot et al., 2019).

Whereas this research (in the form of surveys and interviews) provides a view of the web archiving landscape, ideal workflows for web archival programmes and insights into what web archivists say they are doing, it does not detail how practices are actively enacted and shaped within the organisational contexts within which they are based. By virtue of the methods used, they do not provide in-depth information about the day-to-day decisions, activities and processes that facilitate web archiving in practice. And although work like Truman's (2016) raises much-needed questions about how different professional decision-making practices impact the nature of what is collected and preserved (e.g. between bibliographic and archival approaches to web archiving), the dynamics of how these decisions get made remains understudied.

With an increased focus on facilitating scholarly research using web archives, a number of resources outline both the conceptual and practical challenges of engaging the archived Web. Despite a relative increase in scholarly engagement with web archives (such as the examples provided earlier in this chapter), here the aims have been to address the gap between the widespread collection of web archives and their historically limited use by researchers (Dougherty et al., 2010; Hockx-Yu, 2014a; Meyer, Thomas and Schroeder, 2011). Whilst recent edited volumes (Brügger and Schroeder, 2017) and projects (e.g. the BUDDAH project)¹⁰ have provided much-needed empirical examples of the potential opportunities for research provided by web archives (Schroeder and Brügger, 2017), they have also outlined the many challenges that persist for researchers wishing to understand how collections are amassed in practice. Qualitative work like that from Dougherty and Meyer (2014) provides insights into these challenges through interviews with 'experts' across the field of Internet researchers and archivists/librarians, who demonstrate conflict and concern over issues of provenance, the subjectivity of records and the lack of transparency in technologies used to collect web archives. This point is reiterated by Maemura et al. (2018) who, through a qualitative study of three web archival collections, develop the concept of *web archival provenance* and demonstrate a myriad of intertwined sociotechnical factors that shape both decision-making (around how and what to archive) and the interaction of individual decisions in the creation of web archive collections. And although not focused on scholarly use, Summers and Punzalan (2017) also provide insights into the ways that practitioners are enacting *appraisal* in web archiving (or decisions surrounding identifying and assigning value to resources), as well as the role of bots and automated agents in the construction of collections.

Researchers have also begun to use web archives themselves as a reflexive method for critically analysing the effects of web archiving on the nature of what is archived and made available. Through the use of large-scale analytics to identify and assess 'archival artefacts' and biases (Ben-David and Huurdeman, 2014), these studies have demonstrated the effects of temporal drift in the composite re-presentation of web assemblages in the Wayback Machine (Ainsworth, Nelson and Sompel, 2015), compared platform-specific coverage between the 'live Web' and the Internet Archive's UK Domain Dataset¹¹ (Hale, Blank and Alexander, 2017) and examined the effects of different mechanisms for selecting 'seed' nominations in web archives (Milligan, Ruest and Lin, 2016; Nwala, Weigle and Nelson, 2018). These approaches examine and attempt to 'reverse engineer' (Gehl, 2017) the infrastructural contingencies of web archiving and have further revealed the importance of understanding the direct impact of collection activities on the shape and subsequent use of web archives.

¹⁰The Big UK Domain Data for the Arts and Humanities project (2013-2015) generated a number of case studies, tools and peer-reviewed scholarship using the UK Domain Dataset. More on the project can be found here: <https://web.archive.org/web/20161101045802/http://buddah.projects.history.ac.uk/about/aims-and-objectives/> (visited on 17th Jul. 2019, archived on 1st Nov. 2016)

¹¹<http://data.webarchive.org.uk/opendata/ukwa.ds.2/> (visited 15th Jul. 2019)

As all of these examples have acknowledged, collection and curation decisions surrounding both what is collected and how, as well as the varying degrees to which these decisions are made visible to potential users, have significant impact on users' ability to map the opportunities and limitations of web archives as sources for the past Web (Maemura et al., 2018). However, by virtue of the methods used (and the availability of access), these types of studies are largely focused on using the 'outputs' of web archiving to attempt to remotely re-construct the 'inputs' and processes that enable web archiving in practice.

This thesis complements and extends the studies described above both in methodology and in their aims to address how collection practices structure the nature of future web archival engagement. This research contends that these activities and practices – though often *black boxed*, disregarded or deliberately hidden – are critical for interpreting the affordances of web archives and the types of claims made possible by their use. Here, by employing an ethnographic approach and expanding practice research to include sites outside of libraries and archives, I aim to both widen and deepen the existing field of knowledge surrounding web archiving practices. Below, I outline these aims and the research question addressed in this thesis.

1.4 Aims of the Thesis

Building on the scholarship above, this research re-situates web archives as places of knowledge and cultural production in their own right, by implicating both web archivists and technologies in the shaping of the 'politics of ephemerality'¹² that lead to the creation, maintenance and use of web archives. In short:

RQ: *In what ways do web archival practices (the who, why and how) shape the archived Web?*

Using an ethnographic approach to explore the mechanisms and circumstances surrounding the collection and maintenance of web archives at multiple sites of production, this thesis identifies key practices embedded in the mobilisation, collection and maintenance of web archiving which ultimately shapes the archived Web. Taking an interdisciplinary Web Science approach that draws on STS, critical data and information studies, and archival theory, this research engages with the *performativity of web archiving* and the ways in which web archiving embodies different forms of *knowledge work* that are embedded in the production of web archives. This research examines web archives as socio-material assemblages through the documentation and exploration of web archival practices in the context of three sites engaged with web archiving: the Internet Archive, Archive Team and the Environmental Data & Governance Initiative (EDGI). Of key interest for this thesis is how web archiving is situated within

¹²Taylor (2003) uses this phrase to describe the power and politics involved in the practice of preservation – making the archive the 'liminal zone where objects, files and memories may be lost or retrieved' (Zeitlyn, 2012, p.466).

particular organisational and community contexts, and the ways in which practices are shaped by a variety of sociotechnical factors across this diverse field of practitioners. Drawing on practice theory, the research takes an ethnographic approach – through the use of interviews, non/participant observation and documentary analysis – to document and conceptualise the everyday activities of web archivists.

Here, I want to use the concept of the ‘ethnographic gaze’ to illuminate the aims and contribution of this research to the wider field of knowledge surrounding web archiving. Though rarely defined, the concept of the ethnographic gaze is often used in anthropological scholarship to refer to the object or subject of ethnography with the view of clarifying the role of methods and the instruments of research in the processes of identifying, interpreting and making claims about a field of study. For this thesis, the ethnographic gaze encompasses the ways in which the field of web archiving is brought into focus, expanded and deepened in ways that explicitly draw on the interpretive value of qualitative research to lay bare the subjective and situated actions that embody archiving the Web.

First, I *widen the gaze* to look at sites of web archiving that exist beyond the boundaries of conventional memory institutions. As the Web itself is argued to have disrupted the expectations surrounding who is responsible for archiving the Web, presenting new opportunities for the networked enrolment of new actors, collectives, Web sub-cultures and practices, that are quickly becoming integral to the landscape of web archiving. Second, I *deepen the gaze* beyond standards and best practices to examine the ways that web archiving gets done ‘in situ’. The ethnographic approach therefore enables a deeper examination of the socio-material ways that web archiving is enacted, as well as the meaning-making activities that accompany the ways that web archiving is framed and understood by the participants themselves.

By drawing on the concept of the ethnographic gaze I underscore the ‘situatedness’ of this research to emphasise the ways that the research and findings are still subjective, partial and very much a view from somewhere. The value and limitations of this approach are further discussed in the methodological framing discussed in Chapter 3. Despite drawing both on a range of theoretical framings from archival theory, STS and critical data studies, this thesis is not concerned with best practices, advocating for standards, particular modes or ways of doing web archiving, or defining what counts as ‘legitimate web archiving’. Instead, this thesis takes a broad view of what qualifies as web archiving in an effort to study both the material and symbolic value placed in web archives by the organisations and communities that form the basis for this research.

1.5 Chapter Summaries

Chapter 2, begins by first briefly charting the field of web archiving practices through a presentation of key initiatives and components of web archiving, followed by an examination of some of the known challenges facing web archival practitioners. Here, I further describe and draw on the field of previous research that is relevant to this thesis. The subsequent sections frame the research approach by drawing on a theoretical framework composed of archival theory, critical data studies and STS approaches to the *materiality of knowledge*; before further discussing the contribution of this thesis.

This is followed by a description of the methodological approach in Chapter 3, which includes an examination of the theoretical underpinnings and implementation of the chosen methods, as well as a description of the limitations such a strategy has for understanding web archival practices in the context of the sites selected for investigation (the Internet Archive, Archive Team and the Environmental Data & Governance Initiative [EDGI]). The enactment of the ethnographic approach and methods in each case are detailed, as well as the ways the approach has drawn on thematic analysis (Braun and Clarke, 2006) and facet methodology (Mason, 2011) to frame the analysis. The thematic analysis of the findings from each site is described, with particular significance placed on the chosen facets through which the empirical chapters are presented.

Subsequently, three empirical chapters detail the main findings of this thesis, each through the lens of three facets of web archiving chosen to best represent each site of investigation: *infrastructure*, *culture* and *politics*. Each facet frames the results to answer the research question through emergent themes that shape the ways that web archiving is done. Borrowing from ‘facet methodology’ (Mason, 2011) as an analytical framework, each facet provides one ‘methodological-substantive’ plane or surface view of web archiving practices in each site of investigation.

Chapter 4, examines web archiving at the Internet Archive through the facet of *infrastructure*. In this chapter I conceptualise *web archives as infrastructure* to foreground the wide-ranging implications that these sociotechnical relations have for how web archiving is operationalised at the Internet Archive. The lens of infrastructure enables two observations about web archiving. First, following a relational view of infrastructure, web archiving practices are acknowledged to be situated in space and time. Second, web archiving as infrastructure reveals the heterogeneous labour that enables web archiving in practice. Labour is then conceptualised through four related concepts (knowledge work, translation processes, maintenance and repair) that frame the ways that web archiving is collected and maintained at the Internet Archive.

Chapter 5, examines web archiving through the case of Archive Team, ‘a loose collective of rogue archivists, programmers, writers and loudmouths dedicated to saving

our digital heritage'.¹³ This chapter is centred through the facet of *culture*, providing a dual lens through which to understand web archiving practices as contingent upon the cultural worlds which they create and operate within. Here, *web archiving as culture* reveals the ways that practices – ‘as an observable object for the study of culture’ (Swidler, 2001) – are filtered through and structured by a particular cultural frame that shapes community membership, the nature of how, what and why the Web is archived and the reflexive significance placed on their own web archival activities.

Chapter 6, examines the findings from my work with the Environmental Data & Governance Initiative (EDGI), a collaboration of academics, librarians and civic technology activists formed in the wake of the US Presidential election of Donald Trump to safe guard public environmental data. This chapter is presented through the facet of *politics*, which works to highlight web archiving as politics in several respects. Politics highlight both the centring of web archiving as a tool for ensuring data accessibility but also provides the recognition that web archiving practices themselves are situated, contested and subject to the expertise and political will of a coalition of stakeholders that shape *why* and *how* web archiving gets done. Using the concepts of boundary objects (Star and Griesemer, 2015) and boundary work (Gieryn, 1983), EDGI and the *DataRescue* movement that followed, provides insights into the ways that environmental data became a boundary object that created a solidarity movement and mobilised a diverse set of actors, skills and expertise to participate in political action in anticipation of an ‘anti-science’ administration.

And finally, Chapter 7 reflects on the conclusions of this thesis and presents potential paths for future work. Through a discussion of the cumulative findings from each facet and site of investigation, I reflect on the central research question and aims of this thesis. I further consider the place and significance of each site within the wider landscape of the future of the Web. Prospects for extending this work are also proposed.

¹³<https://www.archiveteam.org/> (visited 26th Jul. 2019)

2

Archiving the Web, Mapping the Field

In Chapter 1, I made the case for the growing role of web archives in the circulation of information and culture online, along with outlining fundamental concerns that emphasised the inherent connections between how the Web is archived, its future use and our understandings of web archives, archivists and the Web. Presented in three main sections, this chapter extends this discussion by further mapping the key initiatives, components and known challenges associated with archiving the Web. Here, previous research, alongside support drawn from archival theory and Science and Technology Studies (STS), frames this thesis and the need for further research on *how* web archiving practices shape the archived Web.

2.1 Web Archiving in Practice

The history of web archiving has been documented to varying extents in existing overviews (Brown, 2006; Brügger, 2011; Webster, 2017) which chart the emergence of a field of practice. Each have used a series of factors to characterise the field over time, including: the tools and technologies used, the frequency and scale of selection and collection methods (e.g. broad versus targeted), and the various motivations behind the creation of web archives. These motivations for web archiving may represent, at least in part, a continuation of classic conceptions of the value and role of libraries and archives as institutions that provide access to cultural heritage, information and knowledge resources; facilitate evidence-based accountability and

promote community memory and identity, amongst others (Barry, 2010; Cook, 2013; Gilliland-Swetland, 2000).

Further background is provided below which examines contemporary knowledge around who is engaged in web archiving activities, the known mechanisms by which web archiving has been implemented, as well as the functions and motivations that drive these programmes.

2.1.1 Key Initiatives

Although most histories of web archiving typically begin with the emergence of the Internet Archive, as Brügger (2011, p.29-30) has pointed out, there were indeed earlier systematic attempts at preserving Internet-based materials. For example, the Electronic Publications Pilot Project (EPPP) is representative of an early phase of Internet/web archiving that focused on the preservation of static web-based e-publications¹ that had the ‘characteristics of traditional publications’ (Electronic Publications Pilot Project Team and Electronic Collections Committee, 1996, p.9). In this way, web archives were positioned to extend and support the traditional functions of library services for maintaining long-term access to scholarly works (a focus of early ‘link rot’ studies). In the university context these functions have more recently been extended to the preservation of unpublished academic outputs and institutional web resources (NDIIPP, 2012), and includes aspirations towards facilitating various enabling services for web-based research using web archives. All of the above continues to be incentives for the creation of university and national library web archiving initiatives.

As Brügger (2011, p.30) notes, these types of library programmes continued in parallel along with the emergence of programmes focused on the wider remit of capturing dynamic web content beyond just those of ‘publication quality’ or produced by the academic community. Early pioneers include the National Library of Australia’s PANDORA project,² and the (Swedish) Royal Library’s *Kulturarw*³ project (both established in 1996) which focused on efforts to archive web content from their respective national domains. Like the EPPP, the *Kulturarw*³ project began collecting other Internet-based sites (not just WWW sites over HTTP) including Usenet and Gopher sites (Arvidson and Lettenström, 1998).

The Internet Archive, a private, non-profit digital library has also been archiving web resources (with the help of Alexa Internet) since 1996. One of the most distinguishing factors (amongst many) that separated the Internet Archive from other initiatives at the time was the broad transnational scope of its collections. Since 2001, the Internet Archive have made their web archives publicly accessible online via the Wayback Machine (with some caveats). Given this scope, the Internet Archive is commonly

¹This project did also include publications that were being served via other protocols over the Internet, including the Gopher protocol.

² <http://pandora.nla.gov.au>

thought to be the largest collection of web archives in the world, having archived 273 billion webpages from over 361 million websites at the time of writing (Goel, 2016c). In an effort to demonstrate the value of web archives, early collaborations between the Internet Archive, the Smithsonian Institution and the Library of Congress gave rise to some of the first event-based web archives collected around the 1996 and 2000 United States Presidential Elections (Kimpton and Ubois, 2006, p.202) and 9/11 (Foot and Schneider, 2010). More on the Internet Archive's web archival practices is further discussed below.

As the domain has expanded over the years, a variety of institutions and alliances have conducted semi-regular surveys with the aim of understanding who is engaged in web archiving and under what conditions and constraints (e.g. Bailey et al., 2014, 2017; Gomes, Miranda and Costa, 2011; Grotke, 2008; Internet Memory Foundation, 2010; National Digital Stewardship Alliance, 2012; Truman, 2016). These surveys provide a high-level overview of the types of organisations that have historically engaged in web archiving, as well as insights into the growth of the web archiving community. In 2011, Gomes, Miranda and Costa (2011) cited 42 web archiving initiatives based in 27 countries archiving web content, including regional and national libraries and archives, non-profit foundations and private sector service providers.³ In 2014, Hockx-Yu lists the International Internet Preservation Consortium (IIPC)'s membership as 48 organisations,⁴ with Gomes and Costa (2014) citing at least 64 web archiving initiatives worldwide as of 2013. However, it is recognised here that these figures are somewhat biased towards the number and types of institutions that were approached for the surveys rather than necessarily a source for definitive statistics on the total number of web archiving initiatives worldwide. Alternatively, the total recorded number of web archiving organisations is greatly augmented by reports by the Internet Archive that over 450 partner organisations use their *Archive-It*⁵ subscription service for web archiving, internationally (Bailey, 2016). According to the Archive-It service, 54% (241 of 509) of their listed partner organisations engaged in web archiving are universities, with almost 20% (101 of 509) composed of varying types of libraries, archives and museums – including state, national and public institutions (Figure 2.1).

More clues to the types of organisations engaged in web archiving (at least in the North American context) are given by the most recent National Digital Stewardship Alliance (NDSA) survey (Bailey et al., 2014). Of the 92 survey respondents, 52% (48)

³One of the outputs of the Gomes, Miranda and Costa (2011) survey was the creation of a Wikipedia page listing existing web archiving initiatives, internationally. This page has since been expanded to include additional information about each initiative, and now includes entries for tools and other web archiving resources: https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives (visited on 4th Apr. 2016)

⁴The IIPC represents a coalition of memory institutions and private-sector organisations established in 2003 to coordinate the preservation of Internet content 'for future generations' (IIPC, 2004). The IIPC collaboration facilitates the creation of standards, policies and tools for web archiving in an effort to 'fulfil the vision of universal coverage of Internet archive collections' (IIPC, 2004).

⁵<http://www.archive-it.org>

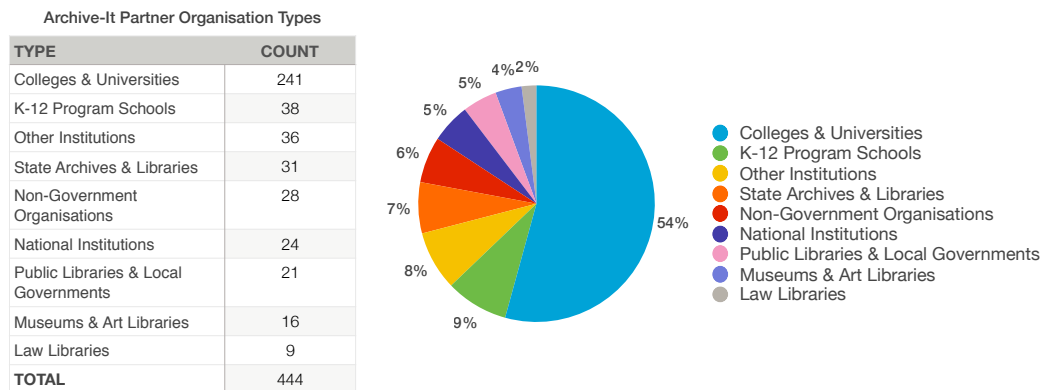


FIGURE 2.1: Archive-It partner organisation types, as tallied from the Archive-it collection website (Archive-It, n.d.) (visited on 10th Jan. 2017).

are based in university or college libraries, with 28% in either archives (14) or state governments (12) (Bailey et al., 2014, p.4). These surveys, in combination with the Archive-It partner organisational types, paint a mixed picture of the web archiving landscape – one that seems to be predominantly composed of university libraries and archives, national libraries and government organisations.

The increasing participation of national libraries and archives in web archiving may be explained by government digital record retention policies and the proliferation of non-print legal deposit schemes for the capture of web-based publications. In several countries web archiving is a legal requirement undertaken by national libraries and archives. For example, in the UK legislation requires The National Archives and the Public Record Office of Northern Ireland to capture records of importance created by the UK government (Pennock, 2013). This has led to the creation of the UK Government Web Archive⁶ which archives the government's web presence. This includes government websites (.gov) and social media content posted using government departmental profiles (National Archives, 2014), as well as functions to redirect users to the web archive in the event web pages and resources are no longer present on the 'live Web' (National Archives, 2011). Additional legislation such as the UK Electronic Communications Act (Pennock, 2013) and other industry-specific requirements for record retention (including web communication and social media) have resulted in the need for web archival expertise in the private sector, as well. This has led to the creation of a number of services, internationally – for example, Hanzo Archives,⁷ MirrorWeb⁸ and Aleph.⁹ These types of services offer web archival selection and collection strategies aimed at regulatory compliance, as well as litigation prevention,

⁶<http://www.nationalarchives.gov.uk/webarchive>

⁷<https://www.hanzo.co>

⁸<https://www.mirrorweb.com>

⁹<http://aleph-archives.com>

‘digital continuity’ and ‘brand protection’ for corporate, financial and legal institutions.

From organisations to collectives, these initiatives are being supplemented by projects and individual researchers actively engaged in the creation and use of web and social media archives for personal use and research purposes – albeit not necessarily with long-term preservation in mind. Mirroring early discourses in the web archiving community,¹⁰ Rogers (2014) observes that only recently has social media transitioned from being considered ‘pointless babble’ (Kelly, 2009) to being recognised as ‘data’, or rather, a legitimate resource for researchers interested in studying large-scale discourse, ‘patterns of social behaviour’ (UK Data Forum, 2013, p.16), global event tracking and the effects of ‘network sociality’ (Wittel, 2001).

The rise in research in this domain combined with the limitations of observing (temporarily stable) live web-based transactions (Uprichard, 2012) has inevitably motivated the creation of archives to enable this kind of research. Schneider and Foot (2004, p.116-117) outline how web archives can support a number of different types of web-based research, including discursive or textual approaches to web content analyses and structural or feature analyses to understand the ‘situatedness’ and interconnectedness of web content. This has been supplemented by more recent work that positions web archives as sources for humanities research agendas, including historiographical studies of the origins of the Internet and WWW, and their subsequent impact on present-day forms of the Web (Brügger, 2013), contemporary politics and web communication patterns (for example, Milligan’s [2016] and Ruest and Milligan’s [2016] work on Canadian elections and political parties over time), and a range of topics as demonstrated by the Big UK Domain Data for the Arts and Humanities (BUDDAH) Project.¹¹

These types of research projects and initiatives are first contingent on the collection of web archives as data resources, whether by individuals, collectives or institutions. As others have noted (and are further described below), the methods by which web archives are collected and maintained have significant implications for how they are used. The following section discusses some of the known methods used for selecting, collecting and managing access to web archives.

2.1.2 Components and Technologies of Web Archiving

Methods for capturing and curating web archives differ across organisations, and indeed reflect the range of motivations for web archiving. Existing overviews (Brown,

¹⁰For a discussion of early debates surrounding the quality of web content and its suitability for preservation see Masanès (2006, p.2-6).

¹¹The various case studies undertaken as part of the BUDDAH Project are available on their website here: <http://buddah.projects.history.ac.uk/2015/07/09/project-case-studies-now-available> and here: <http://buddah.projects.history.ac.uk/2016/04/19/more-project-case-studies-available> (visited on 30th Jan. 2017)

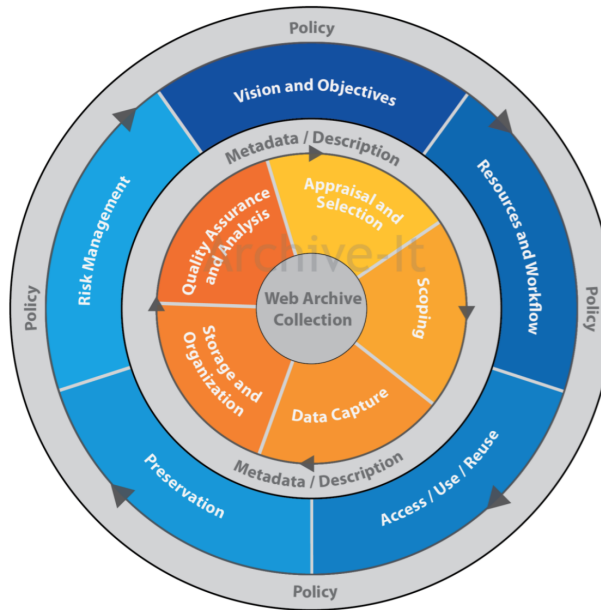


FIGURE 2.2: The *Web Archiving Life Cycle Model* for best practices in web archiving (Bragg and Hanna, 2013, p.3).

2006; Masanès, 2006), literature reviews (Niu, 2012) and best practice documents (Bragg and Hanna, 2013) provide models (Figure 2.2) for mapping the general components of web archival activities. Third-party services, such as those offered by the Internet Archive (Archive-It)¹² and previously by the Internet Memory Foundation (Archivethe.net)¹³ have emerged to facilitate web harvesting, technical expertise and infrastructure to support the collection of web archives for institutions with a range of archival needs. These ‘subscription services’ offer web tools for collecting and curating web archives, including interfaces for managing and initiating seed lists and web crawls, reporting tools for assessing the quality of captures, and forms for adding descriptive metadata to collections. The NDSA described an increase (from 60% to 63%) in the use of external services for web archiving by respondents between the 2011 and 2013 surveys, including 20% of respondents reportedly using in-house tools or some combination (16%) of both for collection (Bailey et al., 2014). In this section, several aspects of practice are broadly considered as they relate to the thesis, including selection, collection and quality assurance, and description and access activities. The role and function of dedicated services for web archiving, as well as other tools used within the field will be explored further as they relate to each component of practice.

¹²<http://archive-it.org>

¹³<https://web.archive.org/web/20180928191235/http://archivethe.net:80/fr/index.php/about/180> (visited on 15th Jul. 2019, archived on 28th Sep. 2018)

Selection and Scoping

Some programmes, such as that of the Library of Congress (Grotke, 2011b) and the National Library of Australia's PANDORA project, collect on the basis of thematic and event-driven selection criteria, or based on web resources chosen by subject specialists for national or local significance. In addition, both initiatives are driven by 'opt-in' approaches to collection methods – in that where possible, permissions to archive are first sought from content producers (National Library of Australia, 2005). The British Library's UK Web Archive¹⁴ utilises a 'hybrid' approach to archiving, building thematic collections of websites pertaining to 'political, cultural, social and economic events of national interest'. The UK Web Archive also uses the UK's non-print legal deposit scheme (UK Parliament, 2013) to legally facilitate semi-automated harvesting of web objects on the basis of the '.uk' country code top-level domain (ccTLD) and other national identifiers (British Library Web Archiving Team, 2014; Hockx-Yu, 2013).

Within the library and archives context, nominations for selection can come from subject specialists or 'recommending officers' who determine themes and/or specify target starting URLs for web crawling – otherwise known as 'seed lists' (Grotke, 2011a). In addition, seed lists are frequently publicly crowdsourced (via blog posts, Twitter and email) as a means for polling significant or popular websites around a certain topic area or event. Examples of this can be seen throughout the last two decades, including the 9/11 web archive (Foot and Schneider, 2010, p.63), as well as the recent End of Term¹⁵ and 2016 Rio Olympics (Byrne, 2016) nominations. Other projects such as those driven by the Archive Team community have used semi-automated 'bots' or agents via Internet Relay Channels (IRC) and Twitter to facilitate archival nominations for specific sites or media – for example, to allow users to nominate videos for preservation from the social media platform, Vine after the announcement that Twitter was shuttering the service (Archive Team, 2016; Vine, 2016).

Collection and Capture

The development of web crawlers, originally created for the purpose of indexing and navigating the Web (rather than preserving it) had a significant impact in propelling and technically enabling the large-scale collection of web objects through the semi-automated traversal of web links and pages (Schneider and Foot, 2008). Large-scale 'client-side' archiving (Masanès, 2006) typically involves the use of 'link-based' web crawlers such as Heritrix¹⁶ or HTTrack¹⁷ to index and download web content over

¹⁴ <http://webarchive.org.uk>

¹⁵ The End of Term project is a collaboration between the Library of Congress, the Internet Archive, the University of North Texas and others, and aims to archive the United States federal government web domain prior to each Presidential administration transition. Further information about the project can be found here: <http://eotarchive.cdlib.org/background.html> (visited on 16th Jul. 2019)

¹⁶ <http://webarchive.jira.com/wiki/display/Heritrix>

¹⁷ <http://httrack.com>

the HTTP protocol (Pennock, 2013). Web crawlers undertake a recursive process of HTML parsing and link extraction, starting from a site domain, host, or seed list of URLs, as provided by the user (Mohr et al., 2004, p.9). The crawl is then determined by a number of parameters (timing, frequency) and seed instructions (including a desired link ‘depth’ and scope), subsequently writing and logging the outputs to chosen file formats – typically in the WARC format, the accepted ISO-standard for web archival storage (ISO, 2009). Increasingly dynamic web technologies have led to evolutions in the development of crawling technologies, including browser-based crawlers such as Umbra (a browser-automation companion tool for the Heritrix crawler)¹⁸ and Brozzler,¹⁹ as well as ‘high fidelity’ web archiving tools such as Webrecorder.²⁰ Browser-based harvesting tools such as Brozzler facilitate web crawling (in this case) by combining the Google Chrome/Chromium browser and warcprox²¹ to open web pages, run a series of ‘behaviours’ based on the platform, screenshot the page and extract ‘outlinks’ before exporting a WARC file.

Whereas document-centric approaches utilise web harvesters, third-party services and legal deposit schemes to facilitate the collection of data; social media archives are often enabled by platform-specific Application Programming Interfaces (APIs), data re-sellers, as well as, in the case of Twitter, the (rare) use of ‘collaborative publisher agreements’ (Hockx-Yu, 2013) for the bulk deposit of data in selected repositories.²² The availability of server/desktop-based and command-line tools for capturing and analysing social media data is also enabling archiving by individual researchers and projects with the skills to wield them. In the case of Twitter, tools such as twarc,²³ TAGs,²⁴ yourTwapperKeeper,²⁵ Twitter Database Server²⁶ and others all offer mechanisms for using various Twitter APIs to download social media data, as well as offer differing analytical capabilities for interacting with the archives. These types of software (many of which are now defunct) and mechanisms for accessing social media

¹⁸ <https://github.com/internetarchive/umbra>

¹⁹ <https://github.com/internetarchive/brozzler>

²⁰ <https://webrecorder.io/>

²¹ <https://github.com/internetarchive/warcprox>

²² One such agreement is that made between Twitter and the Library of Congress in 2010, which stipulated the deposit of public tweets, dating back to the establishment of the company in 2006 (Raymond, 2010). In 2013, the Library of Congress released a white paper (Library of Congress, 2013) on the status of the transfer of tweets in which it gave a high-level description of the technical and economic barriers it had encountered during the (on-going) construction of an infrastructure that would support large-scale archival access to potential researchers. Since this time there has been some backlash to the *Twitter Research Access Project*, during which there have not been any updates from the Library of Congress directly. However, the US Government Accountability Office did issue a report (USGAO, 2015) which reveals that though the project convened a group of stakeholders (the *Twitter Access Group*) to determine the ‘functional requirements to support research access’ to the archive, the Library of Congress expressed significant failings in the construction of strategic plans for budgeting, risk management and scheduling deliverables. This not only offers some insights into the status of the availability of the archive, but also into the challenges social media archives present (at scale) – even for an institution such as the Library of Congress.

²³ <https://github.com/edsu/twarc>

²⁴ <https://tags.hawksey.info>

²⁵ <https://github.com/540co/yourTwapperKeeper>

²⁶ <https://140dev.com/free-twitter-api-source-code-library/twitter-database-server>

data however, are highly dependent on the access provisions granted by the target platforms themselves. As seen in recent years with significant changes and restrictions placed on API-access to platforms such as Facebook, these sorts of provisions are frequently subject to change by platform providers, in turn creating fragile and often unreliable access mechanisms for researchers wishing to study these platforms (Freelon, 2018). These shifting access constraints, in combination with the need to capture more contextual elements surrounding social media data (Walker, 2017b) – beyond the rendering of social media posts as structured data (e.g. spreadsheets and JSON data) – has lead to efforts to combine API collection of social media with the use of conventional crawling technologies. This use of crawlers to capture social media content has further emphasised the challenges of web archiving the dynamic Web – which are presented and discussed below in Section 2.2.

Storage and Access

According to Niu (2012), the degree and manner in which metadata is generated for web archives is constrained by both the scale of the archives and ‘the resources available’ to the organisation. Broad scale archives rely on automated metadata extraction (e.g. status codes, resource size, time of capture, MIME types) and indexing by the crawler and other post-processing tools. In other instances, institutions rely on the creation of manual (in combination with automated) metadata for collection/website/web page description, often using the Dublin Core metadata standard, MARC records or Library of Congress subject headings (Grotke, 2008; Pennock, 2013). Based on data collected by the Archive-It team, Bragg and Hanna (2013, p.20) report that ‘90% of Archive-It partners generate collection level metadata, 60% generate seed metadata, and 15% generate document level metadata’. Gray and Martin (2007) describe a setup where administrators recorded descriptive metadata (‘title, date, subject, description, language, type, format, coverage’), as well as ‘administrative data’ associated with the capture and review process, including the software used to harvest the collection. In the past, the Library of Congress has used a custom tool (DigiBoard) to allow seed nominators to enter metadata pertaining to various aspects of the collection, including subject, language and information about the permissions process at the time of nomination (Grotke and Jones, 2010).

Access to web archives is dictated by a number of factors, both driven by policy, technologies and software that facilitate public interaction with the archived Web. In 2001, the Internet Archive began opening access to web archives through the Wayback Machine (Figure 2.3, Figure 2.4), named after the ‘WABAC Machine’ in the *Peabody’s Improbable Histories* segment on the US cartoon series, *The Rocky and Bullwinkle Show* (Green, 2002; Mieszkowski, 2001). Until recently the query interface for the Wayback Machine remained relatively unchanged, requiring all users to know the exact URL of the website they were looking for in order to gain access to the

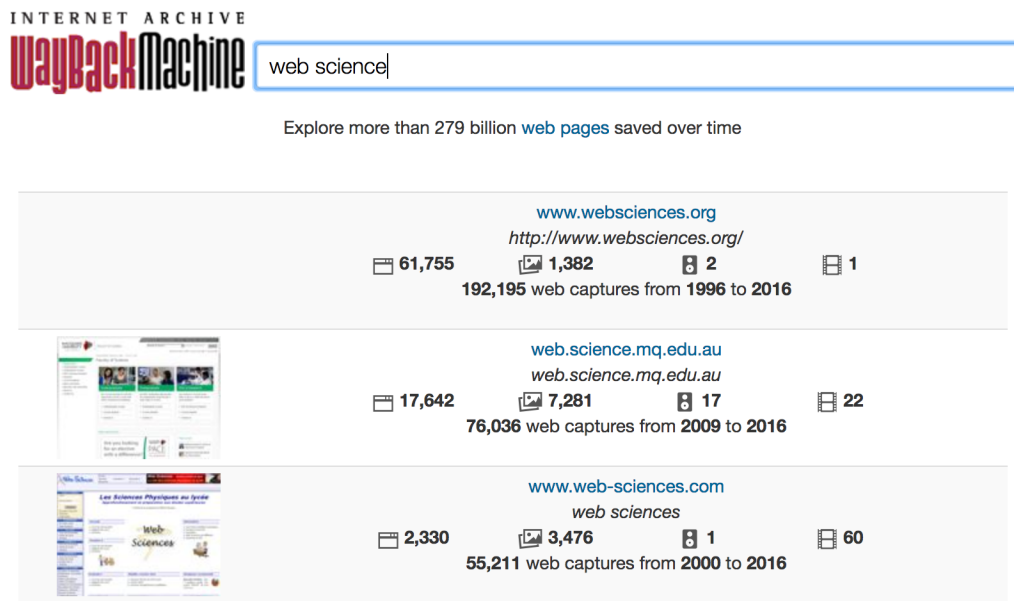


FIGURE 2.3: The Wayback Machine with keyword search.

relevant web archive. This navigation system raised many challenges for researchers (in particular) who have desired a more exploratory approach to accessing archives (See Milligan, 2016, for discussion). In late 2016 however, the Wayback Machine revamped the tool and underlying index and now facilitates predictive text-based search, as well as provides capture summaries with descriptive data including MIME types, number of captures and when captures occurred for result sets (Goel, 2016a). The changes in Wayback Machine present new questions about both the technical processes which underpin search results and the social ramifications of greater access to the largest collection of web archives in the world. These and related ethical questions associated with access are further discussed in Section 2.2.2.

The Wayback software is one of the primary mechanisms for viewing and discovering web archived resources, as evidenced again by Bailey et al. (2014, p.20) who report that 89% (67 of 75) of survey respondents use some version of the Wayback Machine software as their access platform. These numbers may also be biased to Archive-It users as the service also uses a modified Wayback to serve partner captures hosted at the Internet Archive that have been selected to be made public online.²⁷ Some Archive-It partners have chosen to create their own interfaces and landing pages into their web archives, with some developing their own search indices, and integrating their Archive-It collections with web archives held elsewhere (Bragg and Hanna, 2013, p.13-16). In 2005, the Internet Archive released the Wayback code as the Open Source Wayback Machine (OSWB), which was subsequently taken over and led by the IIPC in 2013 as the Open Wayback Project (Hockx-Yu, 2014b). The various

²⁷Bailey et al. (2014) describes that web archive collections in Archive-It are by default 'public' and that these defaults may explain why so many collections are not embargoed.

2.2. Problematising Web Archival Practices

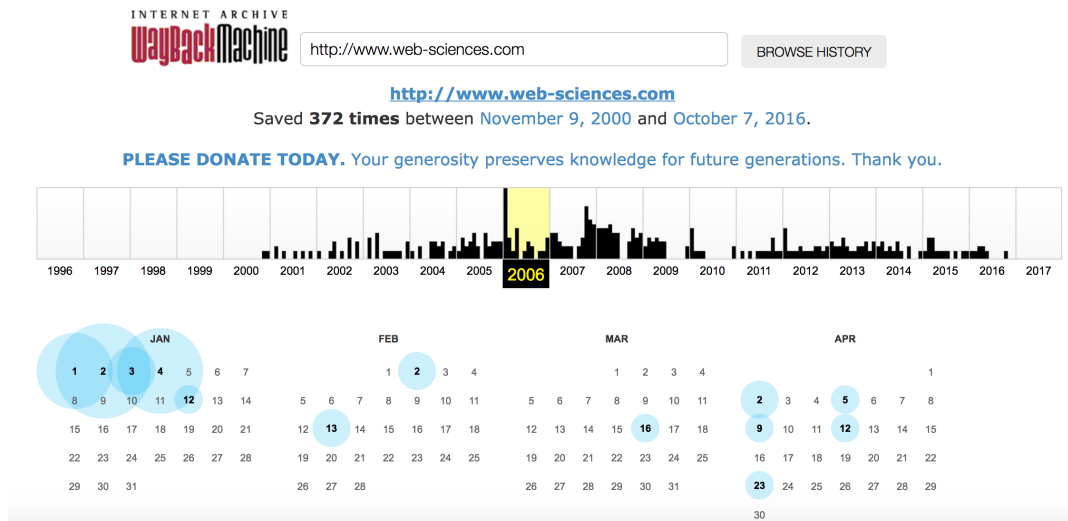


FIGURE 2.4: The Wayback Machine results page indicating the number of ‘snapshots’ that were taken on the day via the size of the circle.

iterations of Wayback represent efforts to expand support and maintenance of the development project to IIPC partner organisations and individuals beyond the Internet Archive. In addition, aggregation tools, through the use of Memento Time Maps, for example, have provided a means for services, such as that of the UK Web Archive²⁸ that provides further access points into web archives held in multiple locations and organisations via the Memento protocol.

2.2 Problematising Web Archival Practices

The field, the practices and technologies used to facilitate the creation and use of web archives are all evolving, but not without their issues. The different legal, economic, technical and ethical challenges to preservation and access presented by both ‘web documents’ and social media have led to a mix of overlapping and divergent collection and access strategies. These include the use of non-print legal deposit schemes, collaborative publisher agreements, platform APIs and web harvesters for capturing online content (Hockx-Yu, 2013). It is recognised here that many of the issues and challenges that have influenced the development of the above practice strategies are intertwined. Nevertheless for the sake of clarity, some known challenges associated with web archiving are discussed separately below in an attempt to begin to explore the ways in which different sets of sociotechnical contingencies shape the nature of what is archived and subsequently accessed.

²⁸<http://www.webarchive.org.uk/mementos/search> (visited on 20th Jan. 2017)

2.2.1 Technological Challenges

Web technologies are evolving at a faster pace than that of web preservation technologies and practice (Dougherty and Meyer, 2014), presenting challenges for preserving an ‘infinite stream with finite resources’ (Leetaru, 2015). The limitations of web harvesters in the face of new markup languages (e.g. HTML5), executable content (e.g. JavaScript and Flash) and other dynamic content (e.g. streamed multimedia, database-driven or password-protected) all lead to missing elements in the representation of web resources in archives (Pennock, 2013). These problems have led to the more recent development of browser-based crawling tools such as Brozzler (discussed above) which attempts to mitigate issues of dynamism by mimicking user ‘behaviours’ with specific (troublesome) platforms and web resources. However, browser-based harvesters present new sets of issues for capture, specifically around the computational memory and overheads required to coordinate capture at scale.

Problems for use are often presented as issues with the ‘quality’ of web archives, as measured by the relative ‘likeness’ between archived web objects and the ‘live web’. As Brügger (2012, p.108) describes, a combination of both collection decisions and ‘technical problems’ leads to archived web objects that are not *copies* of the live web, but rather *contingent constructions* where ‘the process of archiving itself may change what is archived, thus creating something that is not necessarily identical to what was once online’. Recent research by Reyes Ayala (2018) uses an information science approach for measuring quality in web archives, with the aim of improving and easing quality control for web archival practitioners using substantive facets such as correspondence (as a measure of similarity), relevance (pertinence and volume) and archivability to build a theory of ‘information quality’ for web archives. Reyes Ayala also positions the need for grounding information quality issues in ‘human-centred’ approaches to understanding the experience of web archives – beyond metrics which centre the tools and technologies required to render web archives ‘complete’ and representative of its original form. The challenges for scholarly use presented by these quality issues – particularly when web archives are used as evidence for some past state of the Web – are further discussed below in Section 2.2.3.

There is an abundance of research within the field pertaining to improving the efficiency and quality of web crawlers, detecting change in web resources and automating frequency decisions associated with captures (Spaniol et al., 2009), as well as technical overviews of crawling technologies at the time of production (Mohr et al., 2004). Yet, little research exists around the interactive nature and structuring effects of algorithmic and automated agents in decisions around what and when to archive. Recent calls for the study of ‘archival algorithmic systems’ (Summers and Punzalan, 2017, p.831) point towards the need to further consider the performative nature of crawlers and other web archiving technologies, as well as the ways in which different

environmental contexts shape the technological development (Kitchin, 2016) driving collection and access tools.

Helen Hockx-Yu (formerly the Web Archiving Program Manager at the British Library) noted their concern in an interview (Dougherty and Meyer, 2014) that the web archiving domain, particularly from an institutional perspective, is almost wholly dependent on the same software and tools to collect and curate web archives. This includes both the harvesters and curator tools used to manage archives, whether conducted in-house or by commonly used subscription services. The claim is supported by reports that as of 2013, 70% (53 of 75) of surveyed US web archiving institutions use the Archive-It service for their web archiving needs (Bailey et al., 2014, p.18). This raises questions regarding potential path dependencies in the creation and management of web archives, and the subsequent ramifications they may have for the kinds of data that are collected and how they are understood.

The rise and increasing prevalence of social media sites such as Facebook and Twitter virtually ‘hidden from crawlers’ and ‘traditional archiving practices’ (Dougherty and Meyer, 2014) have led to alternative mechanisms for data collection, as mentioned above. Yet tools for social media data archiving are also limited by the technologies and affordances of the infrastructure in which they are trying to access. As social media ‘data access regimes’ are increasingly heavily reliant on the API as the sole form of (bulk) access to platform-based data, epistemological questions are raised. For instance, Twitter’s API rate-limiting and the filtering and spam-limiting behaviour of their search algorithms create concerns over representativeness and reproducibility (Burgess and Bruns, 2012; Driscoll and Walker, 2014), as well as fundamentally implicate the technologies used to create social media archives. Walker (2017b) has charted the methodological implications of decisions surrounding social media collection parameters, as well as the impact of ephemerality on social media research design and findings. All of these examples point towards the fundamental relationship between the technologies of access and the production of archival representations of the Web.

2.2.2 **Legal and Ethical**

Many issues arise from restrictions placed on user-generated content that is collected, stored and used in the absence of consent from its creator. From a legal perspective, legislation often dictates the terms and conditions of data storage and use by collection institutions. For example, in the UK although the Non-Print Legal Deposit laws facilitate the collection of web objects without prior permission from content creators, they also simultaneously restrict the use of web archives to only those who have physical access to Legal Deposit library reading rooms (Hockx-Yu, 2011, 2014a). Here, in compliance with the Legal Deposit statute, the UK Web Archive is only allowed to

provide public, online access to those web archives and collections that are captured with the permission of site owners.

For institutions such as the Danish Royal Library and the State and University Library, who have been archiving the ‘Danish Web’ since 2005 (Schostag and Fønss-Jørgensen, 2012), collection access is limited to researchers and closed to general public use. Schostag and Fønss-Jørgensen (2012, p.117-118) describe the legislative restrictions which present challenges for archival access and use by researchers, including limiting use for ‘scientific purposes’ by ‘researchers at PhD level or higher’. They detail the use of collective licences to overcome issues of copyright, but concede that the archive had yet to satisfy the requirements of the Danish Data Protection Agency with regards to risks to privacy and potential exposure to sensitive personal data which may result from open access to their web archival collections (Schostag and Fønss-Jørgensen, 2012). The British Library and Danish Web Archives present only two examples (of many) where local laws govern and restrict the accessibility of archived web objects.

In the case of social media, the platforms from which web objects are derived provide an additional layer of necessary compliance for collecting organisations. However, Bailey et al. (2014, p.16) report that as of 2013, 75% (59 of 78) of US-based web archiving institutions surveyed did not have an established archival policy in place for social media. In addition to the other obstacles to social media archiving outlined here, another reason for this may be the fluid nature of platform terms and conditions and the challenges this poses for the creation of long-term policies for the collection and storage of social media archives (Lanigan, 2015). For example, Burgess and Bruns (2012) detail how in 2012 Twitter (without community consultation) significantly altered their *Developer Rules of the Road* and *Terms of Use*, restricting high volume (‘firehose’) API access only to their official data re-sellers, forbidding the storage of Twitter datasets with third-party cloud services and open access publication of tweets and secondary data (limiting publication to tweet IDs only), as well as dictating strict terms for the display of tweets.²⁹ In academia, this raises ethical questions regarding both unequal access to data but also methodological concerns over the reporting of research data where personally identifiable information is publicly disclosed.

Ethical concerns have been raised in light of certain technological approaches to consent in web harvesting, including the use of the `robots.txt` protocol (Koster, 1993) as a proxy for consent to crawl any given domain. Thelwall and Stuart (2006) also raise concerns over the potential financial costs incurred by web servers targeted during harvesting routines, as well as the risks to privacy in the cases where the `robots.txt` directives are ignored (as is the option in many software packages).³⁰ Thelwall and Vaughan (2004) consider the possible implications of an unbalanced

²⁹ <https://about.twitter.com/company/display-requirements> (visited on 15th Apr. 2016)

³⁰ For instance, the Heritrix harvester presents users with an option for the crawler to ‘ignore all’ `robots.txt` rules: <https://webarchive.jira.com/wiki/display/Heritrix/Basic+Crawl+Job+Settings> (visited on 14th Apr. 2015)

geographic and country distribution of web-based content in the Internet Archive, framing the findings as an issue of ‘representativeness’ for those using the data for longitudinal or historical research. Others, such as Lor and Britz (2004, 2012) challenge the ‘moral good’ argument for the collection and preservation of representations of cultural production, raising questions over the ethics of primarily western-based (northern hemisphere) institutions archiving the web ‘heritage’ of developing nations. This has been furthered by an increased focus on the ‘right to be forgotten’ in certain spheres and recent studies that have examined the role of copyright and privacy in web archival research practices (Dougherty, 2014).

2.2.3 Defining, Selecting and Using the Object of Collection

Much of the focus of the web archiving community has been on the continued development of technologies and practices for web collection development (Hockx-Yu, 2014a), with an increased attention in recent years on facilitating the scholarly use of web archives (Dougherty and Meyer, 2014; Dougherty et al., 2010; Meyer, Thomas and Schroeder, 2011). This has been supplemented by efforts to provide new tools for scholarly use, including improvements for search systems (Costa and Silva, 2012), web archive aggregation and emulation services (e.g. Memento Time Travel Services³¹ and oldweb.today³²). The increased focus on scholarly use has further highlighted existing divides in the needs of stakeholders – e.g. practitioners (librarians and information science practitioners) and users (humanities and social science researchers) (Dougherty and Meyer, 2014; Dougherty et al., 2010). Through a qualitative study of ‘experts’ involved in web archiving, Dougherty and Meyer (2014) describe a general landscape where scholars (with access to the pre-requisite technical skills) are creating their own web archives for research purposes without the infrastructure to preserve or share them long-term; whilst libraries and archives with preservation capabilities are creating web archives that receive very little use by researchers. Specific ontological and epistemological assumptions made during the collection and curation of web archives are either not made explicit to potential users or are seen as an impediment to their use and/or re-use (Dougherty and Meyer, 2014). Further conflicts are often driven by methodological concerns over provenance, the subjectivity of records and the lack of transparency and metadata for harvester algorithms used to collect web objects.

At the heart of conflicts over the collection and use of web archives are problems with defining the boundaries of the web object to be captured and studied. Whilst recognising that web archives are by nature, inherently and necessarily incomplete (Gomes, Freitas and Silva, 2006), they are also highly ‘subjective reconstructions’ (Brügger, 2008, 2009) of what exists on the live Web at any given time. Appraisal

³¹<http://timetravel.mementoweb.org>

³²<http://oldweb.today>

decisions and selection practices, whether thematic or broad-based harvesting (Mason, 2007; Phillips, 2005), the temporal dimensions of when to capture and for how long (Lyman, 2002) and issues over geographic (Thelwall and Vaughan, 2004) and language (AlSum et al., 2014) coverage and representation of the global Web(s), have all highlighted methodological and sampling concerns over the generalisability of potential research findings based on web archives.

Issues of provenance in web archives – the why, when, and how web archives are collected – have inspired calls for greater documentation around intent, particularly around what to preserve and why (Maemura et al., 2018; Webb, Pearson and Koerbin, 2013). Webb, Pearson and Koerbin (2013) argue that particularly in the realm of digital preservation, institutions have a specific responsibility towards outlining the reasons *why* digital objects need to be preserved before necessarily addressing *how* the individual attributes of those digital objects will be captured and curated long-term. Others have focused on calling for greater transparency in how web archives are built, some with a particular focus on the Internet Archive (Leetaru, 2016). However, practices that surround the capture and maintenance of web archives remain relatively understudied, and for initiatives that sit outside of mainstream memory institutions, continue to exist almost wholly unexamined. Investigations into how curation strategies and collection tools structure the nature of collections - for example, the timing, frequency and length of collection - have the potential to yield insights into the contingencies that lead to the archived Web(s).

Recent qualitative research on web archival appraisal practices by Summers and Punzalan (2017) underlines the value of such an approach for both situating web archiving within wider institutional/archival paradigms, as well as exposing undocumented practices largely missing from the archival record. Summers and Punzalan revealed that practitioners also use a range of tools (e.g. collaborative spreadsheets, web forms) and undocumented practices to manage seed nominations, and found that through the practice of transferring these lists into services such as Archive-It, information about the provenance of these lists remained largely missing from the archival record. Taking a different approach, Milligan, Ruest and Lin (2016) contribute to a discussion of curatorial practices by reverse engineering selection through a comparison of algorithmic, manual and social media-generated web archives associated with the 2015 Canadian Elections. These studies, plus the previously mentioned work of Dougherty and Meyer (2014), reveal the need for additional research on the decision-making processes that surround appraisal and selection in web archiving and the ways in which these under-documented practices present problems related to the use of web archives how they are interpreted and understood.

The decisions to research and archive certain ‘web spheres’ over others have political implications for the formation of historical narratives about the past. Although the preservation of ‘topics of civic and public interest’ are undoubtedly important, questions have been raised about the consequences of a potential absence in web archives

of so-called ‘marginalia’ in digital/web culture(s) (Ankerson, 2012, p.390). Furthermore, the restrictions imposed by access conditions to web and social media archives present additional questions regarding the power to set and influence research agendas. Whereas these types of data offer potential opportunities for new forms of social enquiry (Beer and Taylor, 2013), some have argued for critical engagement with the ramifications of data availability for shaping not only the kinds of questions that can be asked, but also who is allowed to ask and under what conditions (boyd and Crawford, 2012; Manovich, 2012). This has implications for understanding the operation of power in the presence of unequal access to data based on a number of factors including ‘special relationships’ and agreements with commercial platforms and service providers, disparities in funding allocations to support data collection or purchase (Day Thomson, 2016, p.8), and an absence in the technical support required to collect, maintain and analyse complex data sets at scale such as web archives.

2.3 Framing the Materiality of Web Archiving

The previous sections have underscored a number of un/der-documented activities and decision-making processes that matter for how web archives are used and understood. Web archives and associated practices, as sociotechnical phenomena, are structured and organised by an array of actors and environmental factors that actively shape the practice of archiving. This research contends that these (often) undocumented activities and processes are critical for interpreting the affordances of web archives as contingent reconstructions of the previously live Web. Existing surveys have highlighted the prevalence of certain tools and technologies across the field, and though others have pointed towards the presence of potential path dependencies in their use, the ramifications for understanding how web archives are collected and used *in practice* remain unknown. Previous qualitative studies that have engaged in documenting web archival practices have focused on either particular ‘sub-fields’ of practice (including those composed of academic and research libraries) or on certain practices in isolation (e.g. selection or appraisal) of others. Although aspects of web archival practice within the context of Archive-It have been modelled, it is unclear how these components transfer on to the wider domain of web archival practices that exist outside of this subscription service. The power and subjective role of the archivist in web archiving processes remains under-engaged, particularly in relation to their interactions with bots or algorithmic agents.

The argument for a practice-based approach to web archiving is further examined in the next chapter through a theoretical framework. This framework draws on aspects of archival theory and Science and Technology Studies (STS) investigations into the materiality of knowledge. Drawing on Sterne (2014, p.120), here the *material* indicates a focus on the ‘the shape and affordances of the physical world we make and

move through, as well as the constitutive social relations that compose our lived reality'. In various ways – and as will be demonstrated in the empirical chapters – a focus on materiality provides a conceptual tool for understanding the ways that web archiving both *produces* and is *produced by* the dynamic interactions between the 'artefacts, practices and social arrangements' (Lievrouw and Livingstone, 2006) that make up web archives as information communication technologies (ICTs). Taking the recommendation of Lievrouw (2014, p.44), this thesis adopts a framework that 'explicitly accounts for the interplay and mutual shaping of technological tools, human action, and social/cultural formations', enabling an approach to the study of web archiving that reveals and makes visible the sociotechnical work of archiving the Web.

Although this research draws on archival theory, it is recognised too that the differences between *web archives* and other more conventional archival forms may not allow for the simple transfer of one set of theoretical framings to the other. Milligan (2019, p.72) has argued that 'web archives are not traditional archives – not in content, form or conception' in that they rarely conform to the core tenets of contemporary archival practices, like the 'principle of provenance' and the maintenance of 'original order' in archives, or *respect des fonds*.³³ For example, whereas *respect des fonds* emphasises the need for maintaining the original context and order of records, web archives are often thematic collections of disparate web resources on a particular topic, or indeed collected as a result of hyperlinking practices that enable a website or resource to fall within a particular crawling threshold. Web archives are often not given description by archivists or attributed metadata in the conventional sense, leading to vast, sprawling collections of loosely-related material with very little provenance. As such, Milligan (2019, p.67-72) posits that the web archive (as concept) is better aligned with contemporary notions of 'digital collections' or indeed 'back-up' copies of online materials (Owens, 2014) rather than 'archives' in the conventional archival sense.³⁴

On the other hand, whereas the above conceptualisations of archives and record-keeping may reflect particularly constrained notions of the functional nature of what

³³*Respect des fonds* is based on the notion that archivists should preserve the original context in which records were created within the archival record (Ridener, 2009). However, ongoing debates within the archival profession point towards a continued examination of how *respect des fonds* translates in a born-digital context (Bailey, 2013), as well as the ways this principle may in fact privilege narratives of 'naturalness' or objectivity that impede the critical examination of the role of archivists in archival construction and maintenance (Ridener, 2009, p.121). Archival theory and practice is a hundred year-old discipline that deserves far greater space to adequately attend to the nature, origin and evolution of these and other core concepts of contemporary archival practice. Nonetheless, my main assertion here is that these concepts are simultaneously fundamental to archival practice, but still under critical debate and examination by the field (and beyond) in the context of different forms digital archiving and preservation practices.

³⁴Relatedly, others have outlined the ways that 'archive' has been taken up by a range of scholars and practitioners that sit outside the archival profession (Milligan, 2019; Owens, 2014; Theimer, 2012). Whilst I want to avoid a full dissection of the various disciplinary framings here, it bears recognising that these conceptual and practical distinctions have a bearing on how web archives are framed, particularly in light of critiques that they are not 'archival' enough.

counts as *record*, *archive* and *archives* – the ‘records continuum model’ frames recordkeeping over several continuums which require attention be paid to the *processes of archiving* in both theory and practice, not just the end products or individual recordkeeping ‘containers’ (Upward, 1997). As McKemmish, Upward and Reed (2010, p.4448) describe, the records continuum model is a ‘broad and inclusive’ approach to recordkeeping; one which sees it as a form of ‘witnessing, remembering, and forgetting’, rather than confining records to those conceptualisations that originate in paper based systems. The records continuum model therefore enables a focus on a ‘post-custodial’ approach to web archiving that emphasises the ways that practices, people and archives may exist outside the physical walls of memory institutions (Upward, 1997, p.22); to shift the site of enquiry to the places where individual and collective memory are enabled through the activities of leaving, collecting, preserving and representing archival traces. And despite assertions of difference between web archives and other archival forms (and the variability of practices across a range of practitioners inside and outside the library/archives profession), my argument here is that web archiving is fundamentally a form of knowledge/memory production that deserves attention be paid not just to the documentation of practice, but also their theoretical framing in light of decades of critical theory that enables a broader view of what counts as ‘archive’.

Below, I frame my approach by first briefly considering how archival theory has engaged with the ways that practice structures archival knowledge production. These *performative* aspects of archival theory are then considered in light of web archiving, in an effort to make the case for examining the generative capabilities of web archiving.

2.3.1 The Digital Turn in Archives

Postmodernism and its application in archival and social theory lays the groundwork for critically engaging with the material implications of web archiving as a form of knowledge production. Although postmodernism has been argued as heterogeneous in its ‘approach, subject matter and purpose’ (Hardiman, 2009, p.28), emphasis has generally been placed on plurality in ways of knowing, communication as interpretation, a rejection of ‘naturalness’ (Ridener, 2009) and a general ‘incredulity towards metanarrative’ (Lyotard, 1984). There has been a longstanding postmodern interest in fields and disciplines concerned with recordkeeping, through the positioning of records as ‘evidence of process, of activity, [and] of transaction’ (Harris, 2000, p.12). This body of work is the focus here, as it is particularly relevant to an examination of the material social relations and consequences of archival practices.

First coined by Stoler (2002), ‘the archival turn’ denotes a shift from ‘archive as

source' to 'archive as subject', signalling wide-ranging epistemological questions concerning the role of the archive (and the archivist) in shaping and legitimising knowledge and particular ways of knowing. Cook (2001, p.4-5) argues that postmodern archival theory represents a fundamental paradigm shift within a community of practice largely grounded in scientific rationalism, 'archival science', the merits of record stability and the objective role of archivists; towards one which recognises the incompleteness of records, and values context and the interpretive role of archivists in the construction of social memory. This signified a move in theory away from the framing of archives as 'sites of knowledge retrieval' towards a recognition that archives are deeply reflective of and implicated in the production of knowledge (Stoler, 2002, p.90). The conditions of historical narrative-making are intrinsically tied to the processes of archival construction, where certain narratives are privileged and others marginalised through the active reshaping by the archivist (Schwartz and Cook, 2002; Taylor, 2003; Trouillot, 1995). These often invisible 'exclusionary practices' involved in the maintenance of archives, have ramifications for the ways in which archival holdings are often presented as 'being a set of all possible statements' rather than 'the law of what can be said' (Bowker, 2005).

In contrast, Brown and Davis-Brown (1998) focus on the 'technical-rational work' of archivists – or the everyday decisions and practices related to the collection and maintenance of archives. They characterise a profession where the 'explicitly political *who* is often reduced to the technically instrumented *how*' – a sentiment also echoed by critical information studies (Bowker and Star, 1999), as well as practitioners from within the web archiving field (Webb, Pearson and Koerbin, 2013). In light of this, an engagement with the political nature of web archival practice would then include an examination of what comes to count as the 'professional decision-making' (Brown and Davis-Brown, 1998) involved in a host of activities that mark the everyday tasks of archivists based in institutional settings. Brown and Davis-Brown (1998, p.22-29) characterise the different areas as the following:

- **collection development** - includes decisions over what is and isn't to be collected, stored and catalogued in the archive
- **cataloguing and classification** - involves the decisions over the organisation of materials within the archive, as well as the intellectual description practices employed, typically defined and determined by the 'dominant intellectual or political paradigms' through which the materials are viewed
- **circulation and access** - decisions about who sees what and when which are closely shaped by the classification systems employed, and highlights the differences between intellectual vs physical access
- **budget and finance** - who controls the budgets of memory institutions is important for understanding the character and activities of the organisation (funder) itself, as well as the underlying motivations for archiving the materials

- **preservation and conservation** - the degree of care and attention paid to the preservation and conservation of materials in archives ascribes 'value' to records and artefacts. Changes in conservation and/or preservation efforts over time (for example, through format migration or deaccessioning) can be used as evidence for evolving perceptions of value and worth

This form of critical engagement with archival practice facilitates an exploration of archives as socio-political constructions, and enables questions regarding the processes by which 'logical hierarchies' in selection and classification potentially become 'moral hierarchies' (Brown and Davis-Brown, 1998, p.30), with the power to include and exclude. Some of these aspects of practice have been reflected in various best practice and guidance documents (discussed in Section 2.1.2), yet it is unclear how well these types of 'professional decision-making' processes map onto a wider view of what comes to count as web archiving. In particular, very little is known about the practices of organisations and communities that sit outside of traditional memory institutions like national or university-supported libraries and archives, and further, how their practices are influencing the nature of collections held at places like the Internet Archive.

In addition to considering web archives in light of the subjective role of the archivist, the 'digital turn' considers the impact of digital technologies on the archive as a dynamic and contingent information source. Eichhorn (2008) has speculated that the timing of the 'archival turn' coincided with the 'digital turn', describing it as a 'technological and epistemological shift that brought the concept and experience of archives into our everyday lives'. Waterton (2010) describes the archive as a 'protean concept', one which continues to evolve – moving from a 'repository of documents' to encompass 'virtual storage' (Eichhorn, 2008) and information repository technologies of all sorts. The digital turn ascribes the importance of unpicking digital information technologies and not falling into the trappings of either equating them to analogue technologies, nor essentialising their capabilities or potentialities for capturing the cultural record. Cook (2007),³⁵ describes the impact of the transition to electronic archival records and the role of postmodern critique in strengthening the role of the archivist in the digital age:

"We will move from databases to knowledge bases. We will move, in the language of the post-modernists, to re-contextualize our activities: we will reorient ourselves from the content to the context, and from the end result to the original empowering intent, that is, from the artifact (the actual record) to the creating processes behind it, and thus to the actions, programmes, and functions behind those processes. We will move from

³⁵Important to note that this article was first published in *Archives and Manuscripts* 22 (November 1994): 300-328, (now inaccessible online), and was the first instance of the application of a postmodern theoretical framework to archival practice.

nouns to verbs, from records to the acts of recording, from the text to the context behind or through text (or image)” (Cook, 2007, p.410-411).

Here the archival turn is further considered in light of the digital technologies that are intimately tied to both the production and preservation of web resources. In this way, web archives could be examined as a form of ‘dynarchive’, a term used by Ernst (2014) and others (Noordegraaf, 2011) to describe archives that are in a constant state of becoming – or rather, archives that are dynamic, performative and contingent on the material aspects of their production. The concept of the performative web archive is considered in more depth in the following section.

2.3.2 Performative (Web) Archives

Elsewhere, performativity has been proposed as a mechanism for understanding the Web as a sociotechnical assemblage; one which is actively co-constituted and produced through ‘the doing’ (Halford, Pope and Carr, 2010). Here performativity is considered in light of web archival practices, and the ramifications such an approach may have for an understanding of how practice informs the nature of web archival holdings.

One aspect of performativity that is considered here is Butler’s (1990) view on the role of anticipated audiences in shaping action and the ‘naturalisation of practice’. Drawing on Bourdieu (1991), Butler (1996) describes the concept of ‘social magic’, where repeated action leads to patterns of behaviour and belief; a concept which Ladd (2009) interprets in other words: ‘the sedimentation of norms over time transforms the performative into performativity and shifts the locus away from a subject that constitutes an action to a subject constituted by that action’. In this vein, Cook and Schwartz (2002) use the concept of performativity to explore how ‘archival science’ as practice was normalised through the repetition of ‘a sustained set of acts’, and the ramifications for this on the archival record. Cook and Schwartz (2002) argue that postmodern critique of archival practices reveals the supposition – on the part of archivists – that archival ‘audiences’ and users expect objective, neutral archives therefore leading archivists to ‘perform’ the role of ‘objective steward’ in anticipation of this need. Bowker (2005) echoes this sentiment of ‘naturalisation’ whilst also pointing out the role of technologies in the processes that lead to the formation of memory practices:

“What is really interesting is not so much the individual practices and how they articulate a given set of memory practices. Rather, it is how sets of memory practices get articulated into memory regimes, which articulate technologies and practices into relatively historically constant sets of memory practices that permit both the creation of a continuous, useful past, and the transmission sub rosa of information, stories, and practices from our wild, discontinuous, ever-changing past” (Bowker, 2005, p.9).

Waterton (2010, p.653-654) and others make the case for examining the technologies of data archives as integral aspects of the performative, specifically the ‘generative capabilities’ of technologically-enabled data, information, and knowledge which are in an eternal ‘process of becoming’ (Barad, 2003; Haraway, 1988). Whereas others have identified aspects of web archival practice as knowledge production (Schneider, Foot and Wouters, 2009, p.201-213), and have even argued for the ‘democratisation’ of (metadata) description practices in order to enable ‘collective meaning construction’ (Dougherty, 2007), none have been specifically focused on the ways in which these knowledge production practices are both situated and co-constituted by the sociotechnical arrangements of their creation. In this way, web archival practice can be seen as another form of ‘web epistemology’ (Rogers, 2013) that frames what is known. For example, the knowledge production practices that shape the act of making links, connections and ‘cognitive trails’ in databases and other information systems embodies another aspect of the performative (Turnbull, 2003). In web archives, this may be akin to the ordering of web objects in archival formats, as well as through the use of widely accepted metadata and classification schemas to describe and catalogue archival contents. In this example, the decision to use specific classification or metadata schemas – and even the decision to use metadata *at all* – are embedded within particular communities, social worlds, cultural and public spheres that influence and impact upon the nature of archival practices. Unpacking the contingent environments and sociotechnical factors that produce (and are produced by) web archiving practices are the object of this thesis.

In Truman’s (2016) report and study with institutions engaged in web archiving, a question was posed (and answered with more questions) regarding how epistemological differences in the professional practices of libraries versus archives impact the ways in which web resources are first conceptualised and then collected:

“How web archives fit into a larger collection strategy can vary between libraries and archives, but the distinctions between how web archiving is done in each venue is less clear than it is with either physical or other born digital materials. For the most part, librarians use Archive-It and archivists use Archive-It, but the goals for web collecting can vary based on the mission and needs of each institution or unit. If one handles an archived website as a record what are the implications in terms of how it is described, arranged, appraised, stored, and included in planning for long-term preservation?” (Truman, 2016, p.12)

These conceptual distinctions are again echoed in The Library of Congress’ Signal Blog interview with Ben Fino-Radin (a former digital conservator at Rhizome Art-Base)³⁶ who, in an effort to describe the origins behind their job title, distinguishes

³⁶Rhizome is an organisation founded in 1996 that specialises in born-digital art and has been archiving the Web since 1998: <http://rhizome.org>

between a records management and material culture approach to the conservation of born digital (web and Internet-based) art:

"I drew the distinction with my title for two reasons: 1) I am at the service of an institution that lives within a museum, and 2) the digital objects I am cataloging and preserving access to are not 'records' by the archival definition. They are artifacts - and as such require a different kind of care. I am responsible for the stewardship of intellectual entities that are often inseparable from their digital carriers, due to the artist's exploitation of the inherent characteristics of the material. It calls for a high degree of regard for the creator's intent, and a thorough understanding of the subtleties of the materials. A digital archivist tasked with preserving the records of an office probably isn't going to wonder if the use of Comic Sans in the accountant's email signature has artifactual significance" (Owens, 2012).

These questions regarding the impact of prior conceptualisations of significance, usefulness and potential use are manifested in the ways in which digital 'objects' are collected and preserved. In other words, when it comes to archival practices, the social and the material are inextricably linked. Pinch and Henry's (1999) notion of the 'materiality of knowledge' is useful here for considering the ways in which knowledge is also 'embedded in physical artefacts, technologies, and ways of doing things'. This is manifested in how practice produces 'material bodies' (Barad, 2003, p.808-809) but also how materiality is bound to and embedded in practice. In this instance Barad (2003) is referring to how the body (e.g. the anatomy and physiology) actively contributes to the processes of 'materialisation', but in the case of web archives it warrants an examination of how the materiality of technologies (platforms, tools, interfaces, code, algorithms) is both implicated in the production of archives but also potentially produced through practice. For instance, Marres and Weltevrede (2013) reflect on the sociomateriality of 'web scrapers' and the methodological implications for the use of data obtained from harvesters for social science research. Web crawlers – 'automated' agents, bots, algorithms and code – are conceived as not merely passive/objective participants in the collection of web resources, but rather, are intricately implicated in the active shaping of web archives.

The same can also be said for the interfaces that facilitate access to web archives, including (for example) the Internet Archive's Wayback Machine. Others have investigated issues of 'temporal coherence' and the ways in which the Wayback Machine dynamically obscures gaps in the individual capture date-times of web resources in order to present a 'smooth simulation' of composite webpages (Ainsworth, Nelson and Sompel, 2015). These performative aspects of access – the ways in which web archives are both shaped by the collection algorithms and 'replay' tools that dynamically re-present the Web – are necessary for understanding the affordances of web archives. Greater documentation of these processes is required for connecting these

contingent processes with the knowledge work that enables the production and representation of web archives.

2.4 Chapter Summary

This chapter works to re-enforce the need for further research into how web archiving is done. An overview of key initiatives and components of web archiving supports the existence of a long-standing field of engagement with archiving the Web. The development of projects, standards and tools dedicated to supporting web archiving indicates an existing interdisciplinary field of practice, predominantly based in institutional libraries and archives. By drawing on previous research into web archiving, the state of the field was problematised from the perspective of known sociotechnical challenges. I discussed the technical difficulties of capturing a dynamic Web, as well as the current legal constraints and ethical considerations that have had an impact on the nature of what is collected. Issues derived from the problems of selecting and defining the boundaries of web archival collection activities are also discussed in an effort to further frame the ways that web archives are intrinsically subjective representations of the live Web. And yet, the ways in which practices and tools shape the nature of collection and access strategies (e.g. the timing, frequency and length of collection) along with the motivations and meaning-making activities that drive archival practices, remain under-documented. Whereas an abundance of technical research exists around developing tools and improving the efficiency and quality of web archival captures, little research exists around the interactive nature and structuring effects of human, algorithmic and automated agents in decisions around how, what and when to archive (Summers and Punzalan, 2017).

By framing the materiality of web archiving practices from the point of view of archival theory and STS, I emphasise the ways that web archiving technology ‘both embeds and is embedded in social practices, identities, norms, conventions, discourses, instruments, and institutions’ (Jasanoff, 2004, p.3). Here, I underscore the need for research that engages with the ways that web archiving is situated and enacted through heterogeneous systems of values, social practices and technologies that fundamentally shape the nature of what is archived. This type of engagement with performative web archiving exemplifies how web technologies both shape the ‘liveness’ of the archived Web and the ways in which they are implicated in the ‘subjective reconstruction’ of what was once live. It is on this subject that this research is focused.

“There is no dark side in the moon, really. Matter of fact, it’s all dark. The only thing that makes it look light is the sun.”

GERRY O’DRISCOLL, doorman at Abbey Road Studios
(1973, ‘Eclipse’, Pink Floyd’s *Dark Side of the Moon*)

3

Observing Web Archiving: A Methodology

3.1 Methodological Design

This chapter outlines the methodological design taken in this thesis to examine the research question: ***In what ways do web archival practices (the who, why and how) shape the archived Web?***

In addressing this research question, this thesis takes a qualitative approach, involving the use of practice theory and ethnographic methods for engaging with three participant organisations and community groups practicing web archiving. The theoretical orientation and methodological design of the thesis is first discussed, before justifying the use of ethnographic methods across multiple sites. The principles of ethnography are detailed, along with previous examples of the application of ethnographic methods in similar and related fields that have guided this thesis. An outline of the site selection criteria is provided, which led to the selection of three sites: the Internet Archive, Archive Team and the Environmental Data & Governance Initiative. An account of the contribution and execution of each chosen method is then provided, followed by a discussion of the limitations and implications of the research design for understanding web archival practices. And finally, I discuss the use of thematic analysis techniques alongside a ‘facet methodology’ approach to identify and present patterns within the data to address the research question above.

3.1.1 The Qualitative Approach and Paradigm

Qualitative research recognises the complexity and contingent nature of social worlds, requiring in-depth heuristic or descriptive studies of social phenomena (Hammersley, 2012). Qualitative research is embedded within many knowledge production paradigms across the social sciences, from positivist leanings to more social constructivist approaches. This research is rooted in the notion that ‘knowledge is constructed by people in an ongoing fashion as they engage in and make meaning of an activity, experience and phenomenon’ (Merriam and Tisdell, 2016, p.23). This approach prioritises inductive methods to the study of processes, participant meaning and understanding – as well as the use of ‘the researcher as primary instrument’ (Merriam, 2009). In this way, qualitative research is well-suited for the study of phenomena in so-called ‘naturalistic’ or everyday settings (rather than in the controlled or experimental settings often associated with quantitative research), allowing observation of communicative and interaction patterns (Hammersley, 2012) ‘in-situ’.

The mixed qualitative methods approach taken in this thesis complements existing overviews of web archiving, which present a blend of quantitative and qualitative ‘high-level’ data about the state of the web archiving landscape. As discussed in both Chapter 1 and Chapter 2, there have been a number of surveys conducted over the years to assess who is web archiving, and to a lesser extent what they are archiving. This research complements these studies, where the surveys give a sense of breadth when the ethnographic approach may be lacking in aspects of representativeness or generalisability. Furthermore, these studies provide some degree of categorisation for both the types of organisations engaged with web archiving (though limited to the US) and to a lesser extent, the types of web content being archived. This has assisted in the determination of criteria used to select sites of investigation in order to assess, address and document heterogeneity in practices across a field that has historically, had limited research engagement outside of conventional libraries and archives.

The methodological decisions taken in this thesis were focused on the most suitable way of answering the research questions in conjunction and in collaboration with input from participants. It is also worth clarifying that the underlying philosophical orientation of this research supports the viewpoint that in this context, all narratives are partial, particular and situated. The aim of this research is not to ‘know all’, achieve an objective truth, ‘[to get] it right’ (Liamputtong, 2007) or to discover an understanding of web archival practice that already exists somewhere, waiting to be uncovered. This research recognises the subjective and contingent nature of realities, and prioritises the importance of contextualising social activities by both observing and allowing research participants to describe their realities (Baxter and Jack, 2008) first hand within the organisations and communities practicing web archiving. Here it is recognised that meaning is constructed and interpreted, and in the case of the

ethnographic research in particular, I have tried to be explicit and transparent about both my role in the construction of meaning and interpretation throughout.

3.1.2 A Theoretical Framework: Practice Theory

Nadai and Maeder (2005) argue that in contrast to the ethnographic tradition often associated with cultural anthropology, the focus of sociological ethnography frequently spans across multiple groups, places and contexts and thus first requires a 'theoretical framework' to guide the selection of study sites. In the case of this research, *practice theory* provides a theoretical lens through which to engage with web archival practices, as well as a methodological tool for designing and specifying the 'empirical instance[s] of ethnographic fieldwork' (Gómez Cruz and Ardèvol, 2013, p.29).

Practice theory, as 'a set of conceptualisations implying a focus on practice' (Bueger, 2014, p.383), has been used across organisation studies (Nicolini, 2012) and other domains of social science research concerned with accessing the everyday activities of a particular field of knowledge. Central to this approach is the notion that 'human activities [are] centrally organised around shared practical understandings' and that practices are embodied and mediated through the material arrangements of their production and use (Schatzki, 2001, p.11-12). Recognising that there is not one all-encompassing definition for 'practices' across the different theoretical traditions associated with practice theory, this research uses Bueger's (2014) definition which focuses on: action and activities (both mental and physical), 'artefacts and their use' and the tacit knowledge that organises and gives practice meaning. In this tradition, practice theory is thus relevant for:

- An investigation of both physical and discursive activities, as well as the background (implicit) knowledge that gives practice meaning (Bueger, 2014);
- Observing action and artefacts at multiple scales, therefore enabling the detailed site-based study of practices, as well as providing strategies for following and linking practices in the wider field (Warde, 2005);
- An approach that considers the role of materiality and artefacts in practice – for instance, posthumanist practice theory approaches, in particular, take into account the agency of nonhuman and technical agents in the shaping of practice (Schatzki, 2001, p.12).

Practice theory is particularly well suited to the use of ethnographic methods, an approach which has given rise to terminology such as 'praxiography' (Bueger, 2014) to describe the ways in which theory can be empirically applied to the study and observation of practice (*praxis*) in the field of interest. Bueger (2014) argues that praxiography provides a focus for understanding the implicit knowledge that enables and gives meaning to action and activities. Rather than facilitate the ethnographic

study of a ‘whole culture’, a practice theory approach thus allows the focus to be on ‘a cultural understanding of interrelated practices’ (Gómez Cruz and Ardèvol, 2013, p.33-34). In the case of this research, an examination of *practices* signals plurality in a field where there may be many sets of both overlapping and divergent practices that embody different aspects of web archival activities, roles and skills (Postill, 2010, p.16-17). The use of ethnographic methods then enables a situated account of web archival practices within particular sites, as well as the interconnectedness of web archiving within the wider ‘field of practices’:

“The ‘practice approach’ can thus be demarcated as all analyses that (1) develop an account of practices, either the field of practices or some sub-domain thereof (e.g., science), or (2) treat the field of practices as the place to study the nature and transformation of their subject matter” (Schatzki, 2001, p.11).

Three distinct aspects of practice have assisted in the structuring of field observations for this thesis. First, as Ortner (1984, p.149) advises, to examine the individual units that locate and provide *reference points* for understanding particular events and processes. For example, the sites, individual actors, *social types* (e.g. identities, roles), or collective of individuals (e.g. a department or community group) that bounds practices together. This focus can be expanded slightly to also cover what Swidler (2001) points to as *anchors* in practice. Certain individual practices then become an additional unit that may organise and constrain other practices. This facilitated a focus on specific (online/offline) sites where web archiving takes place, but also assisted in the identification of roles, individual activities and artefacts within organisations and communities (for example, crawlers and crawling activities) that organise and define other practices.

A second aspect of the practice approach is to explore the ‘temporal organisation of action’, or the ways in which actions are situated in and relative to various structures tied to time and place (e.g. wider programmes or plans of action within an organisation or community group) (Ortner, 1984, p.150). This sort of approach raises questions around the (temporary) stabilisation of web archiving practices, and has ramifications for an understanding of the relationship between dynamic web resources on the one hand, and the continuous development of tools and expertise to capture and preserve them on the other.

And lastly, the kinds of action (e.g. intentional vs routine or reproduced action) that makes up processes and practices can also structure practice engagement (Ortner, 1984, p.150). In terms of addressing the question of ‘routine’ methodologically, some have advocated for examining ‘moments of rupture’, breakdown or failure in order to explore practice responses or adjustments (Bueger, 2014, p.391). This is akin to exploring some of the known challenges in web archiving that were raised in Chapter 2,

for instance around the breakdown of certain technological contingencies that enable the capture and replay of web archives.

3.1.3 Ethnographic Methods in Practice

In order to address the research question and aims, an ethnographic approach was chosen to document the routine activities of archival practices and the ‘typical patterns of work’ (Hammersley and Atkinson, 2007, p.169) as observed through the problems and solutions which arise in the collection and maintenance of web archives. The research addresses professional and non-professional web archival practices as sociotechnical assemblages with ‘situated meanings’ that differ across organisational settings. Thus, one key practical challenge is a division in the field between sites based in organisational or ‘institutional’ contexts on the one hand, and individual practitioners primarily based online and outside of institutional support for web archiving activities on the other. For this reason, the methodological design of this research is centred upon the use of mixed qualitative methods to observe practice in a variety of multi-site settings, both online and offline, with a focus on documenting the ways in which ‘technologies take on specific social meanings through their embedding within systems of practice’ (Dourish and Bell, 2011, p.74). This research therefore addresses the role of technologies in social practice through the use of observation, interviews and documentary sources.

Principles of Ethnography

Although this research is not what is often referred to as ‘classical ethnography’, it is nevertheless useful to examine the relevant principles of ethnography before outlining the specific implementation and use of ethnographic methods in this thesis.

Ethnography is often used as a descriptor for the *process* of doing research and the end *product* that ‘re-create[s] the shared beliefs, practices, artifacts, folk knowledge, and behaviors of some group of people’ (Pelto, 2013, p.23). Here, ethnography documents the knowledge and value systems that influence action and behaviour with the goal of understanding meaning situated within ‘specific points in time’ (Handwerker, 2001, p.7). With its roots in the field of anthropology, ethnography is best known for enabling insider/outsider studies of ‘otherness’, typically researched through the use of participant and non-participant observation. Emphasis within the field on participant observation has been seen as a rejection of the ‘armchair anthropology’ of the late 19th Century – when anthropologists studied the artefacts of other cultures and constructed theories of faraway social worlds and meanings from a distance (O’Reilly, 2012) – in favour of ‘collecting information about macro processes and wider structures; about institutions, patterns, and norms, as well as about people’s feelings, thoughts and experiences’ (O’Reilly, 2015, p.2). Although ethnography as practice

has evolved, the importance placed on embedded knowledge and contextual understandings of cultural practices and action remains a key characteristic of ethnography in contemporary empirical socio-cultural studies.

Geertz (1973) famously stated that ethnography should not be defined by the methods used but rather conceived as an intellectual, interpretive approach in search of meaning through the production of ‘thick description’ (Ryle, 1968). Here, thick description acts as a mechanism for discerning the ‘many-layered sandwich’ of meaning embedded in everyday interactions (Ryle, 1968), and emphasises the need to carefully parse and interpret material and discursive practices in context. This thesis takes an ethnographic approach to produce detailed descriptive, as well as analytic understandings of how web archival practice works within the context of the chosen field sites. However, an important distinction is made here between ethnography as a *methodology* and the use of an *ethnographic approach*. Due to both the aims and constraints of this research, an ethnographic approach was chosen to facilitate the in-depth study of web archival practices. Whereas classical ethnography tends to focus on exploratory, holistic and longitudinal approaches to understanding individuals and cultural practice, this ‘focused ethnography’ approach is characterised by intense, short-term data collection (Knoblauch, 2005). This type of approach is ideal for organisational ethnographies, which are typically constrained by the nature of access to sites and participants due to the time and economic limitations of the research. Drawing on aspects of ‘mobile’, ‘focused’ (Knoblauch, 2005) and ‘step-in-step-out’ (Madden, 2010) approaches to ethnography, this research relied upon multiple intense, short-term periods of field work to enable detail-rich data collection at each site.

Ethnographic Methods in Libraries and Archives

An ‘upward trend’ has been observed in the use of ethnographic methods in library and information science research, which Khoo, Rozaklis and Hall (2012) attribute to the rapidly changing information technology environment of libraries, the increased need for assessing both the expectations and effectiveness of ‘value added’ services provided by libraries to their respective communities, as well as ‘a growing interest in qualitative analyses of the social lives of libraries’ (Khoo, Rozaklis and Hall, 2012, p.86).

Although there are comparatively fewer examples of ethnographic methods being used within the context of archives, ethnography has been identified as a means for conducting in-depth, comparative and cross-cultural studies of archival practice (Mckemmish and Gilliland, 2013). For the purposes of this research, Gracy’s (2004) ethnographic work within two archives specialising in film preservation provides the methodological stimulus for an ‘archival ethnography’, or rather, an understanding of archival practice *in situ*:

“Archival ethnography is a form of naturalistic inquiry which positions the researcher within an archival environment to gain the cultural perspective of those responsible for the creation, collection, care and use of records” (Gracy, 2001).

Using archival ethnography, Gracy observed and documented the ‘tacit knowledge’ of film preservation practices, or the ‘unstated practices and norms shared among community members’ as they occurred (2004, p.336). The ethnographic approach (using participant observation, in-depth interviews and focus groups) allowed for an examination of the community’s value systems, the ‘shared meanings and disjunctures’ in work practice and the relationships between practitioners with different roles in the archival process (Gracy, 2004, p.338).

Also relevant is Flinn, Stevens and Shepherd’s (2009) ‘community-centric’ approach for examining archival practice through the lens of multiple sites focused on community-driven archives. In an effort to understand the role of community archives in the production of heritage and identity, particularly in relation to ‘mainstream’ memory institutions, they observed archival practices across a number of case studies. The work of Flinn, Stevens and Shepherd holds particular relevance in relation to web archiving activities that sit outside the landscape of conventional memory institutions that are driven by communities with a ‘clear political and cultural mission’ (2009, p.72). Here the use of ethnographic methods presents an opportunity to investigate Archive Team and EDGI, in particular, as centres for *community-based web archiving* in which members collect and control archival materials in and ‘on their own terms’ (Flinn, Stevens and Shepherd, 2009, p.73), rather than as part of a wider programme or set of institutional directives.

Relatedly, the case for examining the ‘situated context of knowledge making practices’ or *epistemic cultures* (Knorr Cetina, 1999) in archival recordkeeping has been made elsewhere, with Ivanov (2017) arguing that ethnographic methods can be used to observe how archival practices are shaped by particular anchoring practices, evaluative logics and sociomaterial genres. Specifically focused on record creation, other studies have used ethnographic observation to understand archives as a form of knowledge production in scientific laboratories (Shankar, 2004); to investigate communication and organisational accountability in recordkeeping (Yakel, 1997, 2001); and to understand both the technical and the social apparatuses that facilitate record creation and maintenance within organisational contexts (Trace, 2002). Observing recordkeeping in law enforcement through the lens of ethnomethodology, Trace (2002) asserts that records are inextricably linked to the social worlds in which they are created, and thus to understand archival practices it is necessary to situate them within the organisational contexts that shape ‘why records are the way they are’.

Although not strictly ethnographic, other qualitative and mixed methods approaches such as interviews, surveys and documentary research have been used in studies of

archives to assess, for example, appraisal practices within university archives (Anderson, 2011); trends and skills in the use of digital methods by archivists (Johare and Masrek, 2011; Kim and Lee, 2009); and the localisation and impact of archival description practices across archives in New Zealand (Battley, 2013). The motivating forces for these studies, as well as their findings, have all emphasised the importance and effect of *context* on the production and management of the archival record and reinforce the value of ethnographic methods for documenting situated practices and the wider interactions between archivists, archival institutions and users of archives.

Observing Practices and Technologies

Ethnographic methods have been argued to enable a more ‘complex’ appreciation of the role and development of technologies in society – beyond a view of technologies as simply ‘functional instruments’ (Prasad, 1997, p.110). One key assumption of this research design is that the direct observation of technologies and their use in web archiving is central to understanding and documenting practices. Suchman (2001) provides relevant assistance here, defining *technology* as ‘the assemblage of skilled practices and associated *logics* characteristic of modern industrial societies’ and *artefacts* as the ‘material production of skilled practice’. This approach allows for an exploration of the materiality of web archiving through a discussion of the relationship between practices and the production of artefacts (both digital and analogue), as well as the role of the environment (the policies, activities, infrastructure, and communities) that actively inform practice.

The aim, however, is not to fetishise web archives as technological objects or material culture, a point mirrored in STS debates which attend to the pitfalls of determinism and advocate for the avoidance of reductionist conceptualisations of either social or technical agents as essentially defined by the other (Suchman, 2001, p.165). In other words, web archives are produced by an assemblage that require attention be paid to both the structure and agency of technical actors, as well as the socio-cultural elements of technical practices – as evidence for the ways in which ‘cultural values are enacted, produced, shared, reified, represented and reaffirmed’ (Dourish and Bell, 2011).

Historically, ethnographic methods have been used in a wide array of studies of the ways in which analogue and digital media technologies are integrated and interpreted through their use, for example through audience or ‘reception studies’. Ethnography has been routinely used in human computer interaction studies and system design (Anderson, 1994) and to understand technological user-requirements and subsequently inform their development (Dourish and Bell, 2011; Suchman, 1985). Ethnographic methods have facilitated so-called ‘internet studies’ to understand different aspects of ‘socialisation’, identity formation and mediated communication practices

in ‘virtual spaces’ (Gómez Cruz and Ardèvol, 2013, p.30). Recognising the development of ‘ethnographic media studies’ is relevant for a study of web archival practices and this thesis in particular, for several reasons.

Web archiving is known to be situated across online/offline spaces, requiring the need to take into account both the role that technologies play in the actual doing of practice and the ways in which practices are mediated through digital communication technologies. Here, a focus on practice enables the thesis to explore the complexities of web archiving as they exist within and across the online/offline (Gómez Cruz and Ardèvol, 2013, p.40). For example, interactions between practitioners in several case studies occurred through online communication mechanisms such as Internet Relay Chat (IRC) and/or Slack.¹ This is of course a matter of necessity for groups such as Archive Team and EDGI, which are online collectives, distributed around the world. However, the use of IRC as a form of communication also presents both ethical challenges for reporting, as well as practical challenges for associating observations with participants with multiple pseudonyms. More on this will be discussed in Section 3.3.4.

As a result, inspiration for designing the online ethnographic component to this thesis was found in classic ‘netnographies’ (e.g. Miller and Slater, 2000), methodological critiques surrounding the importance of participation in observation (Boellstorff, 2012), and research focused on documenting communities that utilise IRC, in particular (Coleman, 2014; Nocera, 2002; Reid, 1996). As part of the ethnographic approach, observing online interactions has been complemented by the use of what Kozinets (2010, p.104-106) calls ‘archival netnographic data’, or previously collected or archived community/communal interactions from online forums. As will be discussed in Section 3.3.2, public archives of Archive Team IRC logs were used to supplement interview and observational data, and provided a rich, longitudinal dataset with which to explore community interactions and practices over time.

3.2 Selecting the Sites of Investigation

Chapter 2 established that there are several different types of organisations and groups known to be archiving the Web. And yet, the specifics about why, what, and how this is accomplished – and more importantly, how these practices and groups shape what is preserved of the Web – remains understudied and in most cases, detailed accounts of practices remain non-existent. Although there would undoubtedly be value in a single, in-depth case study of web archival practices, the choice of multiple sites facilitates the contextualisation of practices within different socially, culturally and technically constructed spaces and places, as well as enables the assemblage

¹Slack (<https://slack.com>) is a group collaboration and communication tool (originally based on the IRC protocol).

of themes across these specific organisations and groups. This research therefore borrows from *multi-site ethnography* (Marcus, 1995) to inform the selection of multiple sites that demonstrate the domains of ‘community life’ – as ecologies, social organisations, developmental cycles and cosmologies (Traweek, 1992, p.7) – through which web archiving is manifested. Here, multi-site ethnography enables the construction of contexts that include and connect multiple sites and places, facilitating both the in-depth study of each site and the ‘making of connections’ between contexts and cases (Morita, 2014, p.222).

A limitation of this approach (as opposed to a large-scale survey, for example) is that the methodology does not necessarily speak to the prevalence of particular practices across the entire population or field (Yin, 2009, p.56). However, the recognition of this limitation also presupposes a definitive boundary or set of variables through which to define the ‘the field’ of web archiving – an *a priori* assumption that this thesis questions and further blurs through the selection of sites that have previously remained under-examined or largely ignored in web archival research. In light of this, I emphasise that this approach does allow for analytical insights into the array of practices used at these three sites in ways that both broadens and deepens an understanding of web archiving, as well as sheds light on sites that are disrupting expectations about both the role and remit of conventional memory institutions and who should be defined as ‘typical’ actors in this space (these points are further discussed below).

Furthermore, the selection of multiple sites (coupled with multiple data collection methods) also has the advantage of enhancing ‘data credibility’ (Baxter and Jack, 2008) and potentially provides more ‘robust and compelling’ findings than a single case study (Herriott and Firestone, 1983). The more sites included, and the greater variation across sites, as Miles, Huberman and Saldana (2014) have argued, the greater strength and stability in the results and interpretation. This has driven the selection of sites that differ considerably from one another (in terms of their cultures, stakeholders, agendas, organisational structures, non-/professional identities, and more); whilst also enabling the exploration of connections across sites (through the networked exchange of people, expertise, tools and technologies). The choice of three sites offers the benefits of breadth whilst still being manageable within the time and resource constraints of this research.

In the language of practice theory, each case can be seen as a ‘site’ or rather, ‘a place composed of practices and material arrangements’ (Bueger, 2014, p.392). Here, the bounds of each site were primarily defined by the physical, organisational and community boundaries that organises their participation in web archiving. The selection approach worked on two levels: first, in the purposeful selection of the sites of investigation and second, in the selection of participants within each site. The processes that drove the selection of participants within each site are discussed as they relate to each method of data collection (described below in Section 3.3). The selection of

sites was driven by a number of considerations, both practical and theoretical. Practically, the selection of sites was constrained by time, money and access, all of which inevitably limited the choice of sites for this research (more on this is discussed in Section 3.3.4).

Given previous research discussed in Chapter 1 and Chapter 2, I established both a need to *deepen* a sociotechnical understanding of how web archiving takes place, as well as *expand* the field of knowledge surrounding how web archiving works beyond conventional memory institutions that are often centred in web archiving discourse. From the perspective of STS, these two goals are premised on the concept of *heterogeneity* and a desire to, as Latour (1987) advocates, ‘follow the actors’. The call to ‘follow the actors’ is often used in STS to draw attention to the non-human actors at play in sociotechnical arrangements (something also heeded in this research design), but in this instance I use it to highlight the ways that particular sites of web archiving are centred whilst others remain unexamined within the field. As Lievrouw reminds us, many STS approaches to materiality (e.g. actor networks [Latour, 2005; Law and Hassard, 1999] and the co-production of science and technology [Jasanoff, 2004]) importantly emphasise:

“[...] the heterogeneity of technology, that is, as a multifaceted and dynamic phenomenon that entails and imbricates not just artifacts, social practices and relationships, and knowledge, but a *variety* of all these elements” (Lievrouw, 2014, p.32, emphasis in original).

This therefore invites both an in-depth investigation of the sociomateriality of web archiving at particular sites, but also the observation that these arrangements will no doubt be variable and contingent upon where you look. Whereas a conventional sampling strategy might be to establish the archetypal or characteristically ‘typical’ organisational types engaged in web archiving, and then choose one from each category, this thesis emphasises the need to consider the ways that existing data on ‘who is web archiving’ is insufficient for capturing the true diversity of sites and actors currently archiving the Web for preservation purposes. Much of the existing research (presented in Chapter 2) concentrates on the nature of practice and technology development in conventional memory institutions, like academic and national libraries and archives. And in the early stages of this research design, I too began from the premise that it was necessary to include a national web archiving initiative as a site of investigation in order to offer the most contrasting data for a comparative analysis with, for example, the Internet Archive or community-based sites of web archiving.² But as Law (1990, p.11) observes, the ‘follow the actors’ slogan ‘reminds us that we tend to reify, naturalise, or simply ignore, what may be important distributions’ within sociotechnical systems. Here, heterogeneity merits the selection of sites that attempt to break down ‘natural’ categories that delineate or default to particular types of sites

²I further explore the practical challenges of establishing access to a national web archiving initiative and the role they played in this site selection decision in Section 3.3.4.

engaged in web archiving, as well as specific tools (e.g. Heritrix) and associated practices often at the centre of research in this space.³

And whilst Chapter 2 established that there is as of yet, a limited understanding of the everyday ways that practices shape web archives even within known initiatives like national or state-based libraries, even less is known about how web archiving works outside of these contexts. For example, the Internet Archive is regularly framed as the biggest web archive in the world, and yet very little is known about the nature of web archiving at this organisation and the ways in which they rely on volunteer labour and expertise to amass their collections. And despite being designated a library by the state of California, the Internet Archive is far from a ‘typical’ case that can be neatly categorised alongside other library-based web archiving initiatives.

In many ways, the Internet Archive can be seen as a disrupter within the digital libraries/preservation landscape. When the Internet Archive arrived on the scene in mid-1996 they were not the only organisation archiving the Web, however, they offered a organisational strategy for web archiving that significantly differed from other records-based approaches that emerged from the archival and library professions at the time.⁴ According to Lyman (2002, p.46-48), Alexa Internet and the Internet Archive took a distinctive approach to web archiving, emphasising the application of computer science to large-scale search and retrieval, as well as underscoring the benefits of a commercial-nonprofit model for funding web archiving collection activities.⁵ Decades later, the Internet Archive now forms a central component of the web archiving landscape through their provision of both large-scale web indexing and preservation services, as well as steering the direction of standards and practice through the development of tools and technologies that are widely used to support web archiving at scale. However, as Ben-David and Amram (2018) argue, the Internet Archive is often projected as a ‘monolithic entity’ in ways that ultimately obscures the diverse sets of sociotechnical arrangements (people, practice and technologies) that enable web archiving to work. And given the significant reliance on the Internet Archive as a source for the archived Web, understanding the ways the Web is archived and maintained in the Wayback Machine is critical for the evaluation of claims being made by those who make use of this vast resource. Given this centrality and with the aim of deconstructing and complicating the ‘lobster trap’ (Karpf, 2012) view of the Wayback Machine, I made the Internet Archive the first site of this research.

In line with the multi-site ethnographic approach, the focus on heterogeneity then led me to ‘follow the actors’ from the initial research at the Internet Archive to two separate community sites (connected to the Internet Archive): Archive Team and the

³There are, of course, analytical implications for these design decisions, and in Chapter 7 I further explore these in light of the contribution of this thesis and potential future work.

⁴See Chapter 2 which describes other key initiatives at the time, including the National Library of Australia’s PANDORA project and the (Swedish) Royal Library’s Kulturarw project.

⁵As will be detailed in Chapter 4, the Internet Archive was set up as the non-profit arm of Alexa Internet, a commercial web indexing service that collected and donated a copy of all sites visited to the Internet Archive (a practice that continues to this day).

Environmental Data & Governance Initiative (EDGI). Formed in 2009 (Scott, 2009a), Archive Team is an online collaboration of volunteer ‘rogue archivists’ who proactively archive websites at scale for depositing at the Internet Archive. Archive Team’s much publicised involvement in partially archiving GeoCities⁶ in 2009 was positioned by Jason Scott (the co-founder) as the event that launched the online collective into ‘Internet consciousness’. References to the ‘ad-hoc collective of guerrilla archivists’ (Milligan, 2017, p.140) are fleeting within the web archiving literature, and are often limited to brief acknowledgements of their contribution to specific collections (like the GeoCities web archive), with very little written about Archive Team in terms of their everyday web archiving practices. However, even at the outset of this research it was clear that as an online, distributed community of volunteers, Archive Team was regularly developing custom technologies and mobilising large-scale campaigns to archive sites and platforms in danger of going offline. As such, the selection of Archive Team as a site of investigation, in particular, enables the ethnographic exploration of how web archiving works both outside the professional expertise and (legal and resourcing) constraints of conventional memory work. In short, what happens when web archiving is under the dynamic direction of a self-described group of ‘rogue archivists, programmers, writers and loudmouths dedicated to saving our digital heritage’?⁷

The third and final site selected was the Environmental Data & Governance Initiative (EDGI), an activist coalition of academics, librarians and archivists, lawyers, civic technology experts (and more) that formed in the wake of the 2016 US Presidential Election. In addition to their on-going advocacy and activism around environmental data justice (Dillon et al., 2017), EDGI and partner organisations orchestrated the ‘DataRescue’ movement in 2017; a series of 50+ public events aimed at ‘rescuing’ and archiving climate change-related information and data hosted on United States government websites. Motivated by widespread concerns that the incoming US administration would remove online access to climate change-related data, DataRescue was an effort to organise subject experts and members of the public around the act of archiving as a tool to ensure access to publicly-funded scientific data. DataRescue was a previously unprecedented collaboration model for web archiving: a broad, networked alliance between activists, civic technologists and libraries and archives to conduct real-time web archiving in community settings. Although EDGI founders and members were in the process of documenting their own work through academic writing, at the time of this research design there had not yet been a specific focus on their web archiving practices.⁸ This research presented the opportunity to document these

⁶GeoCities was a popular, early web platform that enabled everyday users to host their own content within community ‘neighbourhoods’ online (Milligan, 2017). Others have estimated that GeoCities hosted around 38 million web pages before it was taken down by Yahoo in 2009 (Shechmeister, 2009).

⁷<http://www.archiveeam.org/> (visited on 26th Jul. 2019)

⁸There has since been work and publications by EDGI members on ‘activist data archiving’ (Currie, Donovan and Paris, 2018; Currie and Paris, 2018), as well as reflections on the DataRescue events in practice (Walker et al., 2018) that informed the analysis for this research. These and others are further discussed in Chapter 6.

practices and frame the ways that web archiving is being used as a tool for political action, as well as situate this use-case within the wider field of web archiving.

3.3 Data Collection

There are several reasons why the use and combination of mixed qualitative methods was required for this thesis. As in other mixed qualitative and quantitative approaches, the use of different methods in combination compensates where one method might be deficient, reiterating that each method only provides a ‘partial view’ on the event in question (Barbour, 1998). Different methods were used at each stage of the research in order to inform subsequent techniques for approaching the questions at hand. For example, interview data focused and enabled targeted observations (Barbour, 1998) with participants, as well as further contextualised previously collected observational data. Each method thus provided complementary data and insights which informed the other, and in the spirit of flexible iterative research design (a particular strong suit of qualitative methods) (Barbour, 1998; Hammersley, 2012), allowed for the project to adapt and favour methods to suit each context. For these reasons – and taking into account the strengths of each method for observing practice in related research (See Section 3.1.3) – a mix of qualitative methods in the ethnographic tradition were chosen including: interviews, non/participant observation and documentary analysis. The implementation and rationale behind each is described below.

3.3.1 Ethnographic Interviews

Ethnographic interviews are a form of research interview that are distinguished by not only the types of questions that are asked, but also the ways in which the interviews are broken down into ‘ethnographic elements’ or ‘speech acts’ designed to elicit specific types of cultural responses from informants (Spradley, 1979). They are often used in combination with non-/participant observation as a mechanism for developing rapport with informants, as well as to clarify existing ethnographic records (observations) or to focus subsequent observation activities. The interviews conducted as part of this thesis loosely followed Spradley’s (1979) framework for the ethnographic interview and ‘developmental research sequence’. Each interview began with reiterating the research objectives and the purpose of the interview. The interviews took a largely unstructured approach – with a few exceptions described below – using a combination of descriptive, structural and contrasting questions (Spradley, 1979, p.60) in direct response to the answers provided by informants within the context of the interview.

The selection of interviewees followed a hybrid approach that was partially driven by the gatekeepers for each site, as well as snowball sampling initiated by each informant at the time of interview. For example, at the Internet Archive, I first approached Archive-It team web archivists that I had met during a conference and had begun to develop a rapport with. The primary contact also listed the names of informants that would be best suited to the questions that I was asking, favouring those that would provide a longitudinal organisational perspective on the development of practices. In all three sites, this approach was then supplemented by informant-led suggestions which occurred during the context of almost every interview – where informants suggested that I talk to others who could contribute more on particular topics, activities and practices that came up during the course of the interview.

In the case of in-person interviews, all informants were provided with a business card which included a link to the project website. All participants received an email that included a direct link to the Participant Information sheet, as well as the appropriate Consent Form detailing their consent and participation in the project (see Section 3.3.4 for further details regarding ethics). The length of each interview varied, ranging from 20 minutes to 2 hours, all of which were audio recorded. Notes were taken during the course of all interviews, in order to guide the interview, as well as provide extra contextual information for following up with the informant.

In total, 33 interviews were conducted across these three sites. All interviews at the Internet Archive took place in person, barring two which took place on Skype. Most of the interviews took place as one-on-one discussions in various closed-door conference rooms located at the Internet Archive. One interview took place at a local restaurant due to time constraints and the informant's work schedule. Several informants opted to bring along their laptops to the interviews, two just in case there was an 'emergency' and one (at their suggestion) in the event that they needed to demonstrate something. I made the strategic decision to conduct some of the interviews at the desks of informants in the hopes that it would give them the opportunity to 'walk me through' their day-to-day activities on their PCs and show the tools and technologies they were describing. This led to one of the interviews (inadvertently) taking place as a joint interview between myself and two informants whose desks were adjoined.⁹

For EDGI and Archive Team, most interviews and interactions occurred online, predominantly through Skype, IRC, Zoom, Google Hangouts and Slack. During the course of my research I also had the opportunity to meet up with several EDGI members, at which point in-person interviews were conducted. On many occasions I used IRC, Slack and Skype messengers to ask clarifying questions to participants related to observations and previous points that arose during the course of interviews.

⁹The upside to this session was it enabled a really lively discussion between two colleagues who had both similar and divergent things to say about their day-to-day activities. Of course there is a risk in these types of sessions - similar to those observed in focus groups - that may prevent certain informants from participating as much as they would otherwise, or shaping their contribution based on the presence of others.

Descriptive questions made up the bulk of questions asked, which provided a means by which to allow informants to describe the range of day-to-day activities of practitioners, as well as give insight into the background and role of each individual within the context of each site. Different types of descriptive questions were asked in order to simulate the experience of specific sets of tasks and activities (as they arose through the interview), including what Spradley (1979, p.86-88) labels as ‘grand or mini-tour questions’. For example, *"What does a typical day involve?"* or *"Can you take me through the steps involved in [—]?"*. These types of questions were designed to allow the informants to generalise and describe regularly occurring or patterns of activities in lieu of time and access limitations for observing certain tasks. For example, the activities that surround the ‘Wide Crawls’ – as managed by two engineers at the Internet Archive - only occur twice a year and I happened to arrive 48 days after they had initiated the most recent crawl. However, mini-tour questions facilitated useful insights into the tasks and activities involved in Wide Crawls.

Structural questions allowed informants to describe ‘domain knowledge’ specific to web archiving and the site, and included questions about the organisation of web archiving roles and activities within each site (e.g. *"What are the different roles within the team?"* or *"When you say [—], what are the different kinds of tasks involved in that?"*). Contrasting questions were occasionally used to ‘discover dimensions of meaning’ (Spradley, 1979, p.60) behind (locally situated) taxonomies for ordering actors, activities, events and objects (e.g. *"What is the difference between [—] and [—]?"*).

Ethnographic explanations (Spradley, 1979, p.59) were also used within the context of each interview. This involved repeating and reiterating the research remit and focus at the start and throughout each interview, as well as providing additional explanations for why particular types of questions were being asked, or to avoid jarring the informant when the types of questions shifted within the context of the interview. Explanations were also used to further develop rapport with informants, personalise and contextualise the questions and interview process. For instance, in one interview I wanted the informant to know that I found an earlier interaction useful for my research, whilst wanting to get their insights on the development of the project (which was the focus of the previous meeting):

"So, I sat in – you were there when I sat in on the Brozzler dev meeting, which I found really interesting for a number of reasons. One of them being how the processes behind developing new tools for web archiving work. I don't know if you want to say something about that – even just generically or about that specific project and how that's developed over time and what the motivations were?"

As the fieldwork advanced, in some instances, the interviews became more structured with time and focus. As the data collection progressed, observations and interviews

focused the data collection in ways that meant I asked certain questions that I would not have previously (as new knowledge was gained about certain processes and practices). Questions were also inevitably shaped by the roles of certain informants – for instance, the topics of conversation changed in focus with those in management or founding organisational roles – or with engineers versus archivists. Overall this means that although there is overlap in most of the interviews, the interpretive role of both the informant and the interviewer in shaping the content of the interview is that much more prominent in the resulting datasets.

3.3.2 Participant and Non-participant Observation

Generally speaking, the purpose of ethnographic observations is to provide access to ‘practices and actions as they unfold’ (Boellstorff, 2012, p.55), in the form of ‘non-elicited data’ that can allow insights into the implicit and embodied activities that form everyday life. This is predicated on the notion that some actions can not be articulated by participants or ‘insiders’ through other research methods such as interviews. Boellstorff (2012) and others (e.g. Nisbett and Wilson, 1977) have long argued that ‘elicitation methods’ such as interviews can not be a substitute for observation as there are inherent differences and disconnects between *what people do* and *what they say they do*. Furthermore, participants do not always have the perspective or ability to report on all aspects of processes – particularly cognitive ones – that underly the decisions and activities that make up practice (Nisbett and Wilson, 1977; Ormerod et al., 2005). Observation therefore, offers another window into understanding the relationship between meaning and action in everyday practice.

This method involved the observation and creation of ethnographic records describing: *what is done* – action, activities; *what is made and used* – ‘cultural artefacts’; and *what is said* – speech acts, discursive activities (Spradley, 1980, p.10-12), in this case involving both human and non-human actors. Participation in observation can be seen as a sliding scale between non-participation on the one end and what Spradley (1980) describes as ‘complete participation’ on the other. Based on different ‘degrees’ of researcher involvement with people and activities, Spradley (1980, p.58-62) further divides participation levels into passive, moderate, active and complete participation (Figure 3.1). The appropriate level of participant engagement is determined by the types of questions being asked and the access provisions of the site and context under investigation. Each level of participation facilitates different types of ‘insider/outsider’ experiences that inevitably shape the nature of observations.

Although there are no definitive boundaries between each type of participation, this thesis employed varying degrees of participation across all three sites, as well as within different scenarios and activities at each site. Fieldwork was carried out at the Internet Archive in two separate trips totalling four weeks in October - November 2016 and February - March 2017. Participation was often dictated by what was

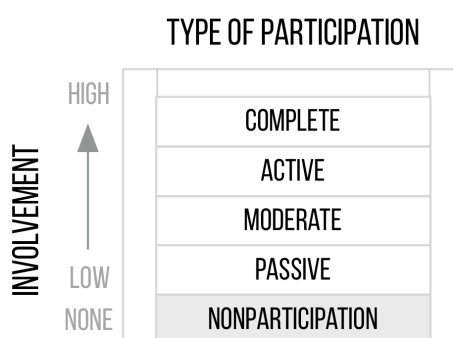


FIGURE 3.1: Model of the levels of participation in observation, as adapted from Spradley (1980, p.58).

deemed appropriate by the host organisation and the ‘event’ or activity in question. For example, at conferences or community events at the Internet Archive I engaged and actively participated in workshops, discussions and brainstorming activities as other staff and invited guests did. Whereas, passive participant observation was used during Internet Archive staff meetings where the influence of my physical presence was enough to rule out ‘non’-participation (even though I did not directly engage or interact with anyone during the course of the meetings).

Participation and observations with Archive Team somewhat waxed and waned with the activities of the collective. In an effort to begin data collection, I began sporadically engaging with Archive Team members from September 2017, through many attempts (some successful and many failed) to make contact with participants for interviews. Under the consent of core Archive Team members (discussed further in Section 3.3.4), during these early stages of my engagement, I used a python BeautifulSoup script to collect historic IRC logs from the #archiveteam and #archiveteam-bs channels hosted by an Archive Team member online. I used these public IRC logs as a form of what Kozinets (2010, p.104-106) calls ‘archival netnographic data’, in order to observe or ‘distance read’ participant interactions over time. This greatly assisted in interpreting Archive Team interactions within the wider context of their activities over time.

Beyond the use of historic logs, participant observation with Archive Team was primarily concentrated during their attempts to archive Tumblr NSFW (‘Not Safe For Work’) blogs between December 2018 - January 2019. I chose this project for close observation as it offered an opportunity to observe and capture the work of Archive Team during a project that required the use of their *Warrior* tool for crowd-sourcing the archiving of large-scale sites. This enabled me to participate, as well as observe the ways that Archive Team enacted web archiving at scale. During this period I followed activities as they occurred on public IRC channels (#archiveteam, #archiveteam-bs and #tumbledown), Archive Team’s GitHub repositories and social media (Tumblr,

Twitter, Reddit). This event was the subject of much of the analysis contained within the second half of Chapter 5.

Participation in the everyday activities of EDGI also evolved over time. Out of all of my sites of enquiry, EDGI involved the greatest degree of embedded participant observation. In many ways, over the course of the data collection, ethnographic observations evolved to embody auto-ethnographic accounts of practice as I became increasingly embedded within the community and the Archiving Working Group, in particular. In addition to the near constant monitoring of relevant Slack channels, between January - June 2018 I participated in almost 70 working group and 'All EDGI' meetings. After being nominated and accepted as an EDGI member in December 2017, I was sent an email outlining in detail, the steps (that I would eventually take) to familiarise myself with the working communication protocols and organisational practices of EDGI. My EDGI interactions spanned several online forums including Slack, Zoom,¹⁰ GitHub and Google Drive (for Google Docs, Sheets and Presentations). I spent most of my time in the Archiving and Web Monitoring channels, split across the EDGI and 'Archivers' Slacks.¹¹ Each of these EDGI field sites had their own localised climate of member participation and interaction, largely reflective of both EDGI protocols and the everyday processes of each working group. The majority of my observations and interactions with EDGI members took place over Slack and Zoom.

In the case of all observations, ethnographic 'records' were made of non/participation activities with the aim of providing the basis for 'thick descriptions' of practice. A combination of both a paper and digital journal were used – where the notebook provided a quick means for noting down activities when the computer was out of reach. Observation pro-forma were not used however, details surrounding the actors (participants), artefacts/objects (technologies, tools) and activities were recorded step-by-step (where possible) to produce ethnographic accounts. In some cases the participant narrated the activity as it happened in realtime which provided a collaborative account of the activity in question. When and where possible I journaled detailed accounts of various online/offline encounters, including preliminary analytic notes regarding my observations and interviews. I also journaled at the start and finish of each day in an effort to capture personal reflections on the time spent at each site. As advocated by Hammersley and Atkinson (2007), I tried to clearly demarcate between my personal reflections, analytic notes and the accounts that arose directly from participants in an effort to clearly and more accurately represent them in the subsequent analysis stages.

¹⁰Zoom is a group online video calling and messaging service which EDGI uses for all of its working group, 'All EDGI' and Steering Committee meetings.

¹¹EDGI engaged with two different Slacks. The EDGI Slack was developed as a 'members only' space in the early stages of the organisation and the other, Archivers, as a public Slack that arose during community DataRescue events for technology-focused volunteers to organise. More on these events will be discussed in Chapter 6.

3.3.3 Documentary and Secondary Sources

Documentary sources in various forms were collected to supplement the observations and interviews. ‘Documents’ is used here to refer to different types of materials (not just textual) produced by organisations, communities and individuals to describe procedures, policies and preferred ways of practicing web archiving. In social science research, documentary sources have been argued to be performative in nature as they can act as both devices for communicating the ideas and policies of those in question, and serve as evidence for their structuring effects on human activity (Prior, 2003). The use of documentary sources here offers insights into aspects of (otherwise) implicit knowledge that underlies practices (Bueger, 2014, p.401), as well as transactionally documents the extent to which certain actors and contributors (that would otherwise be obscured) participate in the shaping of practice. This is the case for example, with the use of Wikis and GitHub as a documentary sources of practice. Here, wiki entries, and GitHub issues and pull-requests record various metadata surrounding each contribution, which enables a view into some of the ways practices are shaped over time and by particular actors.

Existing policy and practice documentation produced (and made available) by each site and organisation were used as additional sources. In the case of both online communities, wikis (publicly available online) and other supplementary documentation (Google Docs, academic publications) were used to supplement and understand the development of web archiving within the community over time. EDGI, in particular produces a considerable amount of documentation pertaining to the everyday protocols, communications and work of maintaining the organisation and web archiving and web monitoring activities. These documentary sources were made available by the organisation, and provided a window into practice documentation going back to the origins of the organisation. Additional sources were found in the ‘commit logs’ for tools on GitHub for all three sites – including the progression of tool development projects, the participant programmers (staff, volunteers and community members) and the frequency of contributions. Furthermore, as has been proposed elsewhere (Kitchin, 2016), the in-line comments available in computer source code were also considered as a mechanism for understanding meaning behind coding decisions, and provided further data to help contextualise technical development practices within the three sites.

3.3.4 Implications of the Approach

Constraints and Limitations of Site Selection and Data Collection Methods

When coupled with ethnographic methods, multiple site-based studies bring with them a set of design and logistical challenges for data collection and analyses that are worthy of some discussion. Additionally, the choice of these specific sites presented

epistemological challenges associated with the collection and comparison of observation data in both online and offline contexts, and with delineating the extents of communities not contained within either institutional or geographic boundaries.

Arranging and gaining access to organisations and people is a common constraint for this type of research and was therefore affected by a number of factors. From the outset I had originally envisioned targeting at least one site based in a national web archive. As described in Chapter 2, national web archiving initiatives were amongst the first to develop practices and technologies for archiving the Web at scale, and have been central actors in the field ever since the National Library of Australia became the first national library to engage in web archiving back in 1996. However, as other research details, despite a widespread recognition of the need to preserve born-digital web content, many libraries and archives engaged in these activities have limited staff and time dedicated to web archiving, particularly as institutions have outsourced the technical components of web archiving to third-party services (Bailey et al., 2014, 2017). This presented some challenges for locating a conventional memory institution that was both performing web archiving in-house and open to the extensive periods of observation integral to this research design.

Furthermore, establishing relationships with these types of national and state-based organisations takes time, and during the course of networking with national web archives and developing the focus of this thesis, I began to identify possible alternative paths for broadening the field of investigation. Using the theoretical framing described in Section 3.2, I looked towards other organisations and communities that have emerged outside the legal deposit mandates that shape practices in most national web archiving contexts. It should be said that the aim of this framing is not to lessen the significance of conventional memory institutions in the field or underestimate the need for further research that addresses these organisations directly; but rather, to underscore the need to attend to previously unexamined archival efforts that occur outside the context of national web archiving.¹² Despite my contention that the facet approach developed in this thesis (and described below in Section 3.4) presents a fruitful avenue for future work on other sites of web archiving, there are of course analytical consequences to the decision not to investigate national web archives here. To address these, in Chapter 7 I further reflect on the implications of this site selection and the potential for future work that investigates and applies the chosen facets to national web archiving initiatives.

Another potential limitation of the site assemblage lies in what may be an implicit North American-centrism underlying the selection of the Internet Archive, Archive Team and EDGI as sites of investigation. Practically speaking, it bears acknowledgement that these particular sites presented some access advantages for myself as a

¹²During the course of this research I did undertake a three month fellowship at the British Library to study the challenges of using the UK Web Archive for social science research. Although this experience did not contribute to the empirical data used in this thesis, it did inevitably inform my understanding of the situated nature of web archiving across these different types of organisations.

native English speaker in that they removed potential language barriers for research presented by other international initiatives, where English may not be the principal language of communication. Although on the surface it may seem that these three organisations and communities stem from and operate in a largely US-context, the reality is more complex. Whilst it is true that the Head Office of the Internet Archive is based in San Francisco, their collection tools and services, and at least some proportion of their staff and volunteer workforce is international in nature. And although the demographics and geographic distribution of Archive Team participants remains unknown, several of the core participants with which I engaged as part of this research are in fact based across the globe. Yet, despite these facts, there are definite limitations in the extent to which practice observations in these contexts translate to generalisations for how web archiving works, in for example, a European national web archive. However, I believe this limitation is alleviated through the use and contribution of the ‘facet’ analytical devices (discussed in Section 3.4) as critical concepts that facilitate the future portability of this empirical work to other possible sites of investigation.

Access to individuals and staff time (in the case of the Internet Archive) was negotiated on a case-by-case basis, where the organisations themselves set reasonable time frames for observation (deemed to be non-intrusive to staff/member time). Allocating enough time to multiple sites was also a consideration in the methodological design, as increasing the amount of time spent on one decreased resources for the remaining sites. Data overload can be another factor in multiple sites (Baxter and Jack, 2008), and whilst linear rather than concurrent data collection was planned to alleviate this aspect of the research, it proved difficult given the flexibility required to undertake the fieldwork for each site. Overall, a balance was struck between the desire to collect more data and the requirements of the research – namely to answer the questions during the time and funding allotted to the PhD project.

The research also presented substantive, practical and ethical challenges for the collection, analysis and interpretation of ethnographic data. Ethnography has long been recognised as a cultural construction (Geertz, 1973), and as a performative, produced and contingent product and form of inquiry (Clifford, 1988). Ethnographic research has been critiqued with regards to the role of subjectivity in the interpretation of ethnographic data for producing results that may be considered too individualistic and personal, or too specific to prevent generalisability (Borman, LeCompte and Goetz, 1986). In the ethnographic approach, as with other forms of qualitative research, the researcher is the main instrument of data collection which requires strategies for ensuring the trustworthiness of the results of ethnographic studies.

As such, the thesis has balanced the need for systematic engagements in the implementation of data collection and interpretation, whilst acknowledging the inevitability (and value) of the role and impact of the investigator in this type of research. And whilst mechanisms for triangulating the results and interpretation were central

to the methodological design of this thesis (including the use of multiple methods), the goal was never to provide a ‘holistic’ or complete view of web archiving at each site.¹³ However, where possible, ‘member checks’ (Merriam, 2009) were used during field work to ensure plausibility and that interpretations ‘rang true’ to the experiences of individual participants. In addition, multiple visits to field sites enabled a concentrated version of ‘step-in-step-out’ methods (Madden, 2010) for making the familiar, unfamiliar; allowing for time and distance to re-evaluate, clarify and re-engage with web archiving practices and participants on site. Nevertheless, here it is recognised that ‘ethnographic truths’ are partial and incomplete (Clifford and Marcus, 1986), where the value and utility of ethnography lies in the discursive, subjective and embedded nature of the methods in practice. Here concepts of the ‘rational detached researcher’ (Rooke, 2009) are challenged and despite the importance of employing techniques for ensuring the trustworthiness of this research, ultimately the results will always be interpretive and subjective.

Access and Developing Rapport

Establishing initial access to these sites was only the first step in this ethnographic work. The development of what has been referred to as ‘entrée’ (Schatzman and Strauss, 1973, p.22) or ‘rapport’ (Spradley, 1979), is a continuous process that presented both methodological challenges and opportunities contingent upon a number of factors that contributed to the development of relationships with informants. Spradley (1979) describes this process as various stages that range from initial apprehension on the part of informants through to eventual acceptance and participation in activities with the researcher (see Figure 3.2). These challenges can differ between so-called ‘insider’ and ‘outsider ethnographies’ (Pollner and Emerson, 1983) and in each of these sites, upon arrival, I felt very much an ‘outsider’. With each initial encounter, I oscillated between the desire to make informants aware of my pre-existing knowledge and expertise in other related domains (my ‘insider’ knowledge) versus my desire to minimise influence over what and how information was relayed to me by informants.

My strategies for negotiating this ‘outsider’ status varied throughout this research. A certain degree of apprehension was observed on the part of informants, particularly around the observations. Challenges associated with the observation of technological use (some of which are described in Section 3.3.4) were somewhat to blame, particularly as I identified and negotiated the specific web archiving activities and circumstances under which I could actually observe practices. The development of rapport with participants was most positively impacted by online and offline social interactions, for example, going for coffee or lunch when I had further opportunities

¹³This is further discussed below in Section 3.4 with regards to decisions surrounding the data analysis.

Apprehension → Exploration → Cooperation → Participation

FIGURE 3.2: The process of rapport development from Spradley (1979, p.79), as exhibited by informants and participants in the ethnographic research process.

to explain my motivations, research topic area and more – as well as through strategically informal discussions and interviews. For the Internet Archive, the timing of the fieldwork (coinciding with several events at the Archive) and a field decision to conduct informal interviews first in some cases (prior to observations) assisted in the development of trust and mutual cooperation between myself and participants. In the case of EDGI, the sheer number of virtual meetings attended enabled the development of a working rapport that quickly eased the awkwardness of the dynamic of ‘being studied’.

For Archive Team, the major limitation in this data collection (and subsequent analysis) was gaining access to people for interviews. Whilst I attempted to make contact in various forums (Reddit, IRC, email), the responses were mixed. I did, however have multiple contacts with an Archive Team organiser who facilitated ‘member checks’ as well as consented to the study. Towards the end of the study I eventually did make contact with other newer participants who were involved in the Tumblr case, which was the focus of an intense period of observation and analysis in Chapter 5. Although any failures of representation are my own, I believe the mixed methods approach taken in this thesis alleviated any limitations presented by a lack of interviews in this particular case.

It is worth reflecting further on the role that I played in the nature of data collection across each method used and site observed. The temporal nature of data collection had an impact on the types of observational data collected across the three sites. For example, although I collected nearly six months of observation data from my time as a participant observer with EDGI, much of the data ultimately used in the thesis focused on data traces and interview reflections on the DataRescue events (which ceased before I began the EDGI case study). In some ways, understanding DataRescue became an archival exercise in combining the wealth of digital traces of the events (e.g. websites, presentations, documents) with the oral histories I collected during interviews. However, (as will become clear in the Chapter 6) in many ways the participant observation gave me access to the everyday ways that EDGI reflected and developed specific organisational responses to the lessons they learned from DataRescue. Observations and my role as a participant (in virtual meetings, working group interactions, Slack), provided rich and valuable context for the interpretation of the other data collected for the thesis.

However, I want to reiterate again that the levels of participant observation varied

across all three sites and each presented different types of limitations for access. Whereas I was highly embedded within the day to day activities of EDGI, observation was much more passive in the case of the Internet Archive and to a certain extent, Archive Team. However, in all cases it was the supplementary ‘liveness’ of observations that enabled a richer understanding of the everyday ways that web archiving shapes the nature of the archived Web. In the case of the Internet Archive, ‘being there’ helped understand the ways that web archiving is built into the organisational context of working at the Internet Archive. But I recognise that as with any form of participatory or qualitative research, my presence undoubtedly shaped the nature of what was said and done in some way. In the case of Archive Team and archiving Tumblr, as the observation data was limited to IRC/logs, there were less opportunities for me to influence the nature of what was being done or said during my observations. There was inevitably some ‘liveness lag’ between participant activities occurring across timezones, and it became evident that there were other non-public communication channels that were restricted to core Archive Team operators. This is all to say that, there were aspects of practice and activities that I did not have access to during the course of this research, but it is my belief that the mixed methods approach (combined with member checks described above) limited the potential impact of uneven access across the sites of investigation.

Ethical Implications of Design

In accordance with the University of Southampton’s research ethics policy, an ERGO application (ERGO/FSHMS/23189) to the Faculty was sought and approved for this research.

Initial consent for the study was obtained through coordination with those in charge of web archiving at each organisation. In the case of Archive Team, ‘organisational’ consent was obtained through a main point of contact which was established through another researcher who had previously interviewed them. For EDGI, a detailed Memorandum of Understanding (MoU) was agreed with the Steering Committee which outlined the nature of my research, as well as a framework for consulting the organisation during the process of my fieldwork. Due to the varied nature of each site, three Participant Information sheets were used, including one for individual participants at the Internet Archive where observations were carried out in person (Appendix A), one for the online community members of Archive Team (Appendix B) and another for EDGI that incorporated the details of the MoU (Appendix C). The Participant Information sheets were used to inform participants of the research project and aims, as well as the details of what their participation entailed. Two types of Consent Forms were used as well, one for the organisations (Appendix D) and one

for individual participants (Appendix E). Organisational/community consent was established to grant access to staff and community members, allowing for their interactions to be observed and recorded. Individual participation in the observations was considered on an opt-out basis, whereas individual consent was obtained for all interviews. In the case of Archive Team specifically, it was impractical to gain prior consent from everyone included in the online IRC logs and wiki data because of their use of pseudonyms, an absence of individual contact details and the sheer number of members (estimated in the thousands). As such, several community ‘gate-keepers’ – i.e. prominent members who moderate community activity – were contacted (via email and in person) to establish organisational consent.

All individual participants were allocated pseudonyms. Ethical concerns are raised regarding representation in ethnographic reporting, particularly for those participants where prior (individual) consent was not obtained (for the reasons stated above). In the case of the Archive Team, IRC channels are synchronous forms of communication (rather than more asynchronous forms of online media, arguably meant for wider distribution or public communication). As such, the data was treated as private communications where individuals – as evidenced by their participation in the channels – may have consented for their posts to be viewed and interacted with by members of the community but may not have anticipated or consented to the future use of the logs for academic purposes. Online pseudonyms are therefore treated as real names, in that they represent the online identities of the individuals who participate in both the IRC channels and other online media outlets (often using the same pseudonym). As such, all reporting of IRC logs in the text uses additional (researcher-allocated, gender neutral) pseudonyms and efforts were made to remove all identifying data points in order to mask participant online identities.

Efforts have been made to manage any sensitive and/or private information that inadvertently arose during the course of the interviews, through the use of pseudonyms and selective reporting. Any information that was deemed sensitive by the organisation from the start of the study (e.g. finances, client information, etc.), was either not included or the organisation in question was consulted prior to the inclusion of the data within the thesis. As was communicated in the Participant Information sheets, full anonymity could not be guaranteed in a small community of web archival practitioners, particularly where some participants are associated with fairly high-profile organisations that form one or more of sites under investigation. This risk was made clear to the participants in both the Participant Information sheets and the Consent Forms. Both the organisational and community Participant Information sheets gave the option for individuals to have their organisational and/or community affiliations revealed in the reporting (i.e. to associate their ‘pseudonymised’ data with the site it came from). The inclusion of identifiable organisations raises ethical issues regarding confidentiality however, the lack of organisational anonymity is balanced here with the ‘minimal probability of harm’ (Kozinets, 2010) to both the individuals and

the organisations. The inclusion of names and affiliations was thus negotiated in each case with the organisations and individuals themselves. All sites of investigation were given the option of not being identified in the findings (with the option of a pseudonym), however all declined.

The issue of anonymity was further complicated by the Internet Archive's desire (at an organisational level) for the materials collected there to be deposited in the Archive. As such, an addendum was added to the individual Consent Forms to allow participants to consent to their interview transcripts being deposited. For each individual participant, I emphasised this addendum and explained that they were under no obligation to either deposit the transcript or indeed participate in the study. It was made clear that they were allowed to withdraw and change their minds at any stage and that they would be given the option to view the transcripts of interviews before they were deposited at the end of the research.

3.4 Thematic Analysis and the Facet Approach

Strategies for synthesising ethnographic data were used to develop analytical themes across the interview and observation data collected. Where practice theory was used to frame and target specific forms of practice (as action, artefacts and knowledge), thematic analysis and facet methodology focused the analysis and interpretation of the results in this thesis. Thematic analysis was chosen as a flexible tool for 'identifying, analysing and reporting patterns (themes)' (Braun and Clarke, 2006, p.79) within each site and across the data collection. Here, thematic analysis offered the opportunity to actively interpret and make connections between the various data collection methods. After initial themes were established through qualitative coding, facet methodology enabled the focused interpretation of particular substantive facets or 'planes' that offered 'strategically illuminating' views into answering the research question (Mason, 2011, p.77).

I used computer assisted qualitative data analysis software (CAQDAS) to organise and transcribe data from across all three sites. I initially used NVivo during the preliminary analysis of the Internet Archive data and then later switched to MAXQDA when NVivo struggled with both the amount of data and the functionality I required for some of the large-scale log data from Archive Team. Using both NVivo and MAXQDA, I transcribed the interview audio recordings as a mechanism for getting familiar with the data (Braun and Clarke, 2006; Reissman, 1993). I transcribed the interview data in a way that attempted to be true to the speaker's intentions through the use of punctuation, whilst still being appropriate for thematic analysis (Edwards, 2001) – as opposed to for example, transcribing for linguistic or discourse analysis.

Qualitative coding was used as a method for assigning categories or themes to individual words, lines and 'incidents' in the data in order to more easily retrieve them

across the dataset (Charmaz, 2006; Merriam and Tisdell, 2016). All data for each site were analysed together in order to first identify common themes present within the site. ‘Open coding’, or the use of quick short codes to describe the interview data (Merriam and Tisdell, 2016, pp.204-205) was first employed as a mechanism for getting to grips with the breadth of the dataset within each site. Lists and groupings were compiled by repeatedly listening to the audio recordings, reading the transcripts and observation records and comparing the subject matter across the data. Paper-based observation records and IRC logs were digitised and imported in an ad-hoc fashion when particular events or observations reflected themes identified elsewhere in the data. Particular attention was paid to ‘identifying moments’ in the data, or when people identified with their practice and the practice of others in order to frame their own experiences – often as right or wrong, informed or not (Charmaz, 2006). Identifying moments were used as a ‘sensitising concept’, or in other words, a ‘way of seeing, organising and understanding experience’ and a jumping off point for further thematic analysis across the data collection (Charmaz, 2003, p.259).

Early analysis also involved the identification and coding of ‘things informants know’ in an effort to elicit everyday practices through the various kinds of participant knowledge (e.g. knowledge about crawl behaviours, scoping rules, reporting tools) and their connections to ways of doing things (e.g. maintenance and quality assurance tasks, de-duplication techniques) (Spradley, 1979, 1980). This provided a device for connecting the (often) implicit knowledge behind practice with discursive and non-discursive practices. This technique was also an attempt to introduce order to a largely unstructured data set. The methods used provided the opportunity for iteratively mapping heterogeneous data and highlighted particular groupings of practitioner knowledge and activities into broad themes.

The final analysis presented here borrows from separate but complementary interpretive approaches advocated by Morgan (1997) and Mason (2011) that use metaphor as a tactic for prioritising different angles to research problems. In *Images of Organization*, Morgan (1997) outlines eight different metaphors, or theories of organisation and management – including the image of organisations as machines, organisms, brains, cultures, political systems, ‘psychic prisons’, ‘logics of change’ and ‘instruments of domination’. Here, metaphor offers a strategy for generating ‘complementary and competing insights’ into the nature and theorisation of organisation (Morgan, 1997, p.6). Each metaphor, Morgan argues, provides distinctive, yet partial perspectives from which to view the multi-dimensional inner workings of organisations. Relatedly, Mason’s (2011) *facet methodology* uses a cut gemstone as a visual metaphor for investigating a field of research (Figure 3.3). In this metaphor, the gemstone (‘the field’) is composed of facets, or ‘methodological-substantive planes and surfaces’, that come in different shapes and sizes and accordingly, have the ability to refract light and illuminate the field in ‘a variety of ways that help to define the overall object of concern’ (Mason, 2011, p.77).



FIGURE 3.3: A cut gemstone and its facets.

Building on the initial thematic analysis and borrowing from the above approaches to metaphor, in this thesis I propose and examine three facets of web archiving: **infrastructure**, **culture** and **politics**. This particular assemblage is reflective of a strategic decision to address the materiality of web archiving, where each facet works to answer the research question by emphasising the dynamic and mutually constitutive interplay of technologies, action and socio-cultural relations (Lievrouw, 2014). I therefore chose three facets (one in each site) to illuminate the material ways that web archiving practice *produces* and is *produced by* the networked relations of sociotechnical labour (infrastructure), particular cultural worlds and negotiated systems of meaning (culture), and the interplay of situated and contested forms of political will and expertise (politics).

It should be said that these particular facets were arrived at through an iterative approach to identifying the key paradigmatic concepts that would offer the most explanatory power at each site during the analysis. These facets, in particular, provided ‘umbrella concepts’ that enabled me to zoom out from the specific themes and practice logics at each site in order to make connections across the field of study. As discussed in Section 3.3.4, the multi-site ethnographic approach produced a wealth of data that presented challenges for distilling heterogeneous and semi-/unstructured data within the time and space limitations of a thesis. In order to facilitate a rich, in-depth discussion of practices and sufficiently address the research question, I chose to focus the presentation of each site analysis through a single facet. And whilst each facet could have been applied to each of the sites (for example, I could have framed the analysis of Archive Team practices as infrastructure, or EDGI practices as culture), I argue that this particular pairing of sites and facets offers critical windows into practice at these particular sites. In this way, the facets also have a cumulative effect in that this approach allowed for the layering of interpretation of practice between sites, whilst enabling a detailed analysis of each individual site through the lens of a particular facet.

Here, I want to briefly acknowledge the risks and limitations of this analytical approach as they relate to both the selection of these specific facets and their application to each particular site. As both Morgan and Mason caution, the use of metaphors (and by extension, facets) can be paradoxical in that each facet has the power to produce views of the field that are simultaneously revealing and distorted. Where particular aspects of web archiving are strategically illuminated, others are also removed from view. However, this is an inherent risk in all research design. By choosing particular methods, analytical approaches and ontological frameworks, researchers make choices about how to best answer research questions with the tools at hand. And whilst I want to avoid over-framing these particular facets as ‘naturally occurring’, it is important to underscore that I chose to emphasise these facets at these particular sites because they offered the most explanatory power for how web archiving is done at these sites and across the field.¹⁴ As Mason puts it:

“The aim of our facet methodology approach is to *create a strategically illuminating set of facets in relation to specific research concerns and questions*: not a random set, or an eclectic set, or a representative set, or a total set” (Mason, 2011, p.77, emphasis in the original).

Collectively, I chose these three facets to directly address the question of how web archiving shapes the archived Web. In this sense, the facets and sites presented here are not a random or representative sample of web archiving practices; rather, they provide analytical devices to tell web archiving ‘practice stories’ (O’Reilly, 2015) that both reveal specific practices at these sites and gesture to the ways they might be applied to the wider field of web archiving. My argument is therefore that web archiving as infrastructure, culture and politics applies to *any* site of web archiving and to demonstrate this contribution, in Chapter 7, I explore the ways that these three facets offer a lens through which to explore the other three sites investigated in this thesis.

3.5 Chapter Summary

This chapter set out to describe the design and implementation of a methodological approach to answering the research question: ***In what ways do web archival practices (the who, why and how) shape the archived Web?***. To address this question, I began by justifying the use of a qualitative approach that draws on practice theory to frame the nature of web archiving as practice (artefacts, action and knowledge). I then outlined the use of mixed qualitative methods for the study of practice, first by building on other examples of their use in libraries and archives, as well as in the observation of technological practice. Here, practice theory and ethnographic

¹⁴As Braun and Clarke (2006) warns, there is a danger in framing an analytical account of themes (or in this case facets, as well) as ‘emerging’ or ‘discovered’, as this denies the active and subjective role of the researcher in identifying and choosing particular patterns and themes to report on (Taylor and Ussher, 2001).

methods offer mechanisms for documenting and situating practices within different contexts of knowledge production, as well as ways of targeting the material aspects of sociotechnical practices.

Building on the case set out in previous chapters, I then outlined the reasons behind the selection of the sites of investigation. A focus on the STS concept of heterogeneity recognises the diverse material arrangements both within and across the field of practice. I argued that given the central role of the Internet Archive in web archiving collection, research and tool development, there is a critical need to document the inner workings of this significant resource. In line with a multi-site ethnographic approach, I then ‘followed the actors’ to two separate community-based initiatives that were chosen to shed light on the ways that web archiving is enacted outside of institutional settings. Here, the field is expanded beyond the boundaries of previous research to include the largely unknown web archiving practices of activist organisations working to archive the Web to their own ends.

I then justified and described the data collection methods used for non/participant observation, and the collection of ethnographic interviews and documentary sources across the three chosen sites between 2016 and 2018. I argued the merits of a mixed methods approach for offering complementary insights into web archiving practices at each site. The varying degrees of participation and ‘embeddedness’ in site-based activities was described, emphasising the ways that observations were subject to the different types of access I was granted at each site.

The implications of the approach were discussed as they relate to the practical constraints and ethical implications of the decisions made during the design and implementation of this research. I acknowledged the value and constraints of the interpretive approach, as well as steps that were taken to triangulate this analysis. Limitations in establishing contact with both individuals and organisations (like conventional libraries and archives) were discussed, as well as the ways that my subsequent experience at the British Library offered additional perspective and insights (if not empirical evidence) into the research topic and field. I also described the steps taken to ensure the ethical integrity of this research project including a range of strategies, such as the use of informed consent (where possible), pseudonyms in all reporting and the removal of information believed to be sensitive to participants.

And finally, I discussed the thematic analysis and facet methodology approach to presenting the results of this research. Here, I outlined the use of thematic analysis techniques and ‘open coding’ to initially identify themes surrounding the nature of web archiving practice in the interviews, observation and documentary data at each site. I then made the case for drawing on facet methodology as an analytical frame for strategically illuminating three facets of web archiving: infrastructure, culture and politics.

The three empirical chapters that follow are presented through each of the facets briefly described above. In this way, each site and facet offers a cumulative view of web archiving across the three sites. The next chapter examines *web archiving as infrastructure* through the case of the Internet Archive.

“[...] perhaps it is time for infrastructuralism. Its fascination is for the basic, the boring, the mundane, and all the mischievous work done behind the scenes. It is a doctrine for environments and small differences, of strait gates and the needle’s eye, of things not understood that stand under our worlds.”

JOHN DURHAM PETERS (2015, *The Marvelous Clouds: Toward a Philosophy of Elemental Media*)

4

Web Archiving as Infrastructure: The Internet Archive

4.1 Introduction

This chapter examines *web archiving as infrastructure* through the case of the largest web archive in the world, the Internet Archive. As a conceptual device, *infrastructure* can invoke a vision of the basic facilities and structures that support societies in everyday life – for example, road and train networks that enable transportation of people and goods, power grids that deliver electricity from source to use and telecommunication networks that invisibly connect people and information. The study of infrastructure has long been a focus of STS scholarship concerned with moving beyond this view of infrastructure as ‘tubes and wires’ (Bowker et al., 2010) towards a relational account that attends to the wider sets of embedded sociotechnical relations and arrangements that enable infrastructures but are often invisible, undervalued and subject to change (Jewett and Kling, 1991; Star and Ruhleder, 1996). *Information infrastructures*, in particular, have been the subject of scholarship that has illuminated the ways that infrastructural practices, processes, classifications and standards, software, data and people shape the nature of how information is produced and circulated (Bowker and Star, 1999; Bowker et al., 2010; Star, 1999; Star and Ruhleder, 1996). Here, I want to similarly complicate web archives and the practices that sustain them.

Information infrastructures are often defined by their fundamental ability to transport information from one place to another, to create access and ‘reach beyond a single location’ (Finn, 2018, p.7). Here, a relational notion of infrastructure extends this to consider the ways that infrastructure also comes into being through the local relations that create and make infrastructure through practice. As Star and Ruhleder (1996, p.381) argue, ‘infrastructure occurs when the tension between local and global is resolved’. This chapter considers this tension between local and global in the context of web archiving and the Web. Here, *web archiving as infrastructure* enables two interconnected observations about the ways that web archiving shapes the archived Web. First, web archiving is enabled by a dynamic set of spatially and temporally situated practices that facilitate the transportation and transformation of web resources from the Web into web archives. Taking a relational view of infrastructure, these practices are enabled through the doing of web archiving, or as I propose, the heterogeneous labour of human and non-human agents that produce knowledge about what, when and how to archive the Web. And whilst in this chapter I primarily focus on the ‘local’ practices of the Internet Archive, a second observation points to the ways these localised practices are realised through networked relations that extend beyond the Archive. Here, as efforts are enacted by the Internet Archive to further embed web archives within the infrastructure of the Web, fundamental questions are raised about the power and implications of web archival labour for shaping the circulation of information and culture online.

Infrastructure as Place/Infrastructure as Practice

The first observation builds on the premise that web archives are a form of *information infrastructure* (Bowker et al., 2010) that are enabled through the networked relations of people, practice and technologies that are simultaneously situated in and produced by particular spatial and temporal arrangements. As such, I begin this chapter by first situating the material arrangements of the Internet Archive as a ‘view from somewhere’ by briefly locating the emergence of the organisation within a particular time and place of significance in the transition to Internet-based electronic publication.¹ Here, the Archive and their web archiving practices emerged in the context of Silicon Valley during the so-called ‘dot-com boom’, enabling the creation of a simultaneously local and global infrastructure for archiving the Web. Early sociotechnical imaginaries for the Web as an electronic library (Stefik, 1997) coupled with the desire to build the so-called ‘Library of Alexandria 2.0’, locates the Internet Archive within a network of efforts to pursue the provision and preservation of net-based information access. I then situate web archiving at the Internet Archive through

¹The ‘view from somewhere’ is a play on wide-spread STS critiques of the representation of knowledge-making practices as objective truth-making, or ‘views from nowhere’ (Nagel, 1986). Further discussion of the significance and ‘particularities of place in knowledge-making’ can be found in Shapin’s (1998, n.4) discussion of the ways that science is both locally produced and travels through practice.

their transition to a community-based ‘third place’ (Oldenburg, 1989), or a free and open community library space. Here, the digital library infrastructure is materialised as ‘bricks and mortar’ (Srinivasan, 2016) within a contemporary setting designed to communicate the grandness of the task of archiving the World’s knowledge.

The remainder of this chapter is focused on the ways that web archiving as infrastructure is enabled and sustained through what I have called *web archival labour* (Ogden, Halford and Carr, 2017). Moving beyond the bricks and mortar version of the Internet Archive, I discuss the ways that web archiving is also enabled by the networked labour of a range of actors who shape the nature of what and how the Web is archived. By extending Downey’s (2014) concept of *information labor*, I argue that web archiving is sustained by dynamic forms of human and algorithmic labour that ‘[enables] and [constrains] the constant circulation of information’ (2014, p.141). The concept of web archival labour is explored through four interconnected components of practice that highlight the ways in which the archived Web is being shaped: *knowledge work*, *translation processes*, *maintenance* and *repair*. Each aspect of labour highlights the ways that web archiving is embedded within particular sociotechnical relations and arrangements of people, practice and technologies that shape how the Web is archived at the Internet Archive. The second observation frames these locally enacted practices in consultation and interaction with a global network of stakeholders that are generated through the production of the archived Web. Here, as web archiving draws on selection techniques that utilise ‘the crowd’ and web archives become embedded into other platforms like Wikipedia, Twitter and other social media platforms, the situated practices of web archiving become temporarily stabilised, reified and even more critical to understand.²

4.2 Locating Infrastructure, Framing the Internet Archive

This section is fundamentally concerned with the ways that web archives as information infrastructures are tied to particular places and time. Taking a cue from Downey (2014, p.148), this section relies on the observation that information labour, or the work to make things work, ‘always takes place in, and depends on, a particular spatial/temporal and political/economic context’. Whereas the subsequent section explores this work in practice (Section 4.3), this section locates web archiving practices at the Internet Archive within the historical and contemporary context of San Francisco and Silicon Valley, California. Below, I briefly discuss the history of the founding of the Internet Archive and some of the ways the organisation has become a digital library with a mission to archive the Web.

²For example, in a recent paper by Zannettou et al. (2018), the use of web archives in right-wing forums contained on platforms like Reddit, 4-chan and Gab illustrates some of the ways that the Internet Archive and other web archives are being embedded in the circulation of information beyond the Wayback Machine. For further discussion of Wikipedia, see Section 4.3.3.

4.2.1 Brewster Goes West and the Dot-com Boom

Following the explosion of Internet usage in the United States in the 1980s and early 1990s (Abbate, 1999, p.181), the invention of the World Wide Web and subsequent release of the Mosaic browser in 1993 (Berners-Lee, Fischetti and Dertouzos, 2000, p.75), and the hype surrounding the future of the ‘information superhighway’ championed by the Clinton/Gore administration (Ankerson, 2018, pp.42-43) – late 1990s Silicon Valley, California became the epicentre of the so-called ‘dot-com boom’.³ This boom came in the form of billions of dollars in venture capital and a particular focus on the development of technologies suited for net-based electronic publishing and discovery. It was during this time that in 1996, Brewster Kahle and Bruce Gilliat established the Internet Archive as an operational non-profit alongside *Alexa Internet*, a commercial web indexing service (Kimpton and Ubois, 2006).

Alexa Internet, in Kahle’s words, was an attempt to ‘catalogue the Web’ and to ‘make it so you knew where you were and where you might want to go next’ online (Livingston, 2007, p.275). Their navigational approach to search and discovery, (akin to a modern day recommender system) ran counter to other tactics emerging at the time (e.g. search engines) as it used a browser-based toolbar to track web pages as users visited them.⁴ The anonymised ‘usage trails’ of the Alexa toolbar stored a copy of the web page in the Internet Archive for posterity. Since the late 1970s, computational search and discovery of information had been a longtime pursuit of Kahle’s, who has (in interviews here and elsewhere) framed the task of building ‘the great library’ as ‘always the goal’ (Livingston, 2007, p.273) despite lacking a number of fundamental components that had yet to be invented. In an interview, Kahle reflects on the goal of networking the world’s knowledge:

“So I wanted to try to get everything online. And that was back in 1980. But there were some things missing—computers, networks, software—so we just started cranking through them. And trying to get that ‘Library of Congress on your desktop’ was sort of how it was talked about in the 80s. I think the Library of Alexandria, version two is probably a better way to look at it—or has Raj Reddy put it, ‘universal access to all knowledge’. And it was to make a smart machine. To make a global brain.”

Kahle, a former MIT student of the famous cognitive scientist, Marvin Minsky, spent the early 1980s working as a lead engineer at *Thinking Machines*, a pioneer start-up in the development of artificial intelligence and networked high performance computing. As a mechanism for connecting the super computers being developed at Thinking Machines, Kahle developed WAIS (Wide Area Information Servers) which became the first Internet-based distributed search and document retrieval system (Kahle et

³Ankerson (2018, pp.25-55) neatly charts this transition period from Internet to Web, marking the significant milestones that can be attributed to the dot-com boom.

⁴Kahle has admitted that they were ‘wrong, just wrong’ about the scalability of search engines like Google (Livingston, 2007; Swisher, 2017).

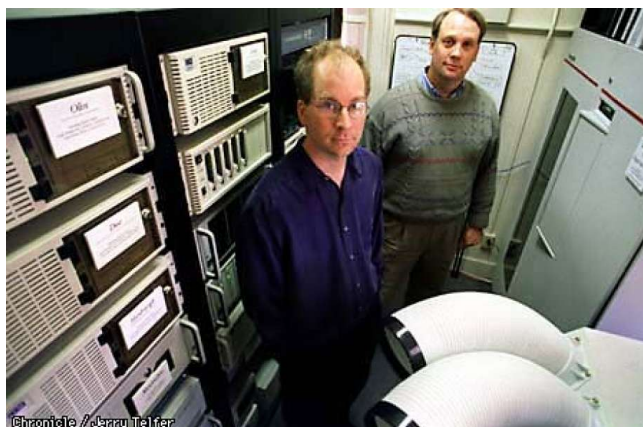


FIGURE 4.1: Brewster Kahle (left) and Bruce Gilliat (right) in 1998 at Alexa Internet's headquarters in Building 37, the Presidio, San Francisco. Image: (Said, 1998)

al., 1992). WAIS was a precursor to the World Wide Web which Tim Berners-Lee released to the public just two years later. To demonstrate the networking capabilities of WAIS, Kahle re-located to San Francisco in 1989 to work on projects with Apple, DOW Jones and others, going on to eventually incorporate *WAIS Inc.* in 1992 (Livingston, 2007). When I later asked why the Internet Archive was based in San Francisco, Kahle relayed the advice given to him by friend and mentor, Bill Dunn when he asked 'where should we put the Internet?'. Dunn advised them to 'go someplace where people don't think you're crazy', and according to Kahle, San Francisco was the 'centre of the universe'.

And so, with the subsequent sale of WAIS Inc. to America Online (AOL) in 1995 for USD\$15 million (Nollinger, 1995), Kahle and Gilliat went on to build the next component in the quest for getting the world's knowledge online: Alexa Internet and the Internet Archive. Having worked together previously at WAIS Inc. and as described in Livingston (2007, p.275), Kahle characterised Gilliat as the 'business-oriented' half of the partnership, whereas Kahle offered the 'visionary side' (Figure 4.1). After Alexa Internet was acquired by Amazon in 1999 for USD\$250 million in stock options (Feuerstein, 1999), Kahle and Gilliat co-directed Alexa as a separate company before Kahle moved next door to the Internet Archive full-time in 2002 – which at the time 'had nobody working [there]' (Livingston, 2007, p.277).⁵

With the expressed goal of making it 'part of the infrastructure of the Internet' (Said, 1998), Alexa was named after the famous universal Library of Alexandria amassed by the Ptoleamic in ancient Egypt. For years, Alexa Internet would be the main source of crawl data for the Archive, a practice that continues to this day (and will be discussed further in Section 4.3).⁶ To demonstrate the new-found goals of the Internet Archive and encourage other libraries to enter the fray, the Archive partnered

⁵Gilliat later left Alexa Internet in 2007 (Conroy, 2008).

⁶This practice was written into the contract when Alexa Internet was sold to Amazon (Feuerstein, 1999).

with the Smithsonian and later the Library of Congress to capture the US Presidential elections in 1996 and 2000, which was followed by more event-based and ‘thematic crawling’ by the Archive in the period to follow.

Despite the dot-com bubble bursting and the US stock market collapsing in 2001, Kimpton and Ubois (2006) describe the years between 1999 to 2002 to be a particularly booming period for the Archive, where they in fact reportedly benefited from a surplus of out-of-work engineers in the Bay Area. Boosted by an expansion in their content collection activities, the development of the ARC standard and a new crawler at Alexa, the Archive moved to make the archived Web available to the public for the first time over the browser (Kimpton and Ubois, 2006, pp.205-207). Developed by Alexa programmers, the Wayback Machine was launched in late 2001, and became an application that Kahle had envisioned other libraries would use to create a global network of other Wayback Machines:

“[...] in 2002, I moved over from [Alexa Internet] to the [Internet Archive] and started archiving away there. But we didn’t have any money, or didn’t feel like we had very much money, and we were getting this stream from Alexa Internet. So we kind of didn’t need to work on that quite as much. But we worked with these national libraries to try to get them to change their laws to not suck and to have them start to play a role. Well, they changed their laws to suck. And they said they were the only ones able to do it. And they weren’t. They didn’t feel free to go and make things available. So they didn’t go and make it so that there would be a lot of different Wayback Machines out there that are publicly on the web. Even though we’d been doing this for 10 years and showing it’s just not a problem. Come on, guys, do it. So we were trying to be a leader and they didn’t follow, which is sort of disappointing.”

Here, Kahle refers to the efforts of the Internet Archive to work with national libraries to create a network of Wayback Machines in the early to late 2000s. These efforts resulted in the creation of the International Internet Preservation Consortium (IIPC) in 2003, and to further collaboration on issues surrounding web archiving standards, practice and tool development which continue to this day.⁷ The frustrations exhibited by Kahle concerning the legal constraints and fears surrounding the remit of national libraries to collect and make these materials publicly available can be found elsewhere in the field.⁸ However, in the context of this research – and whilst wanting to avoid an argument that champions some form of American exceptionalism – the point I want to make here is that the Internet Archive emerged from a space and time that has been intrinsically informed by the social, technical, economic and legal contexts that have enabled them to be risk takers in the field of web archiving. Their practices have

⁷<http://netpreserve.org/about-us/> (visited on 29th Jul. 2019)

⁸For example, Winters (forthcoming) provides a recent summary of the opportunities and challenges of working within the constraints of the UK legal deposit laws.

materialised within a collaborative network of particular sets of expertise rooted in venture capitalism, high tech computing innovation, information access, digital rights and free speech advocacy. These connections and relations will further be reflected on in the context of discussing the labour of web archiving in subsequent sections.

The above section has worked to illustrate the ways that web archiving at the Internet Archive has emerged within the context of the Silicon Valley in the late 90s, early 2000s. To further frame the practices discussed in subsequent sections, below I pivot to the ways that the Internet Archive also acts as a physical community space in San Francisco.

4.2.2 The ‘Spiritual Centre’ of Silicon Valley

Elsewhere the Archive has been proposed as the ‘spiritual centre’ of Silicon Valley (Quinn, 2019), perhaps (at least in part) because the main offices are located inside what was originally constructed to be the Fourth Church of Christ (Scientist). The Head Office of the Internet Archive is situated in the Richmond District of San Francisco, at the intersection of Funston Avenue and Clement, a bustling market street lined with coffee shops, book merchants, Asian markets and cafes. The Archive relocated here in 2009 (Figure 4.2) from their previous location in the Presidio because – as Kahle has been quoted as saying (and as I was told during the weekly tour of the Archive) ‘the building matched [the] logo’. Further insights into the relocation of the Archive to the former church building can be gleaned from video footage taken by staff shortly after the building was purchased.⁹ The videos provide insights into a vision for community reading rooms for the Open Library, spaces for their table-top scribes for scanning books and public-facing window displays for library collections.¹⁰ Here, the Internet Archive is simultaneously rooted in conventional notions of the library as ‘third place’ (Srinivasan, 2016), or an open and free community place, and the infrastructural imaginary conjured up by the vision to archive the world’s knowledge in digital form.¹¹

The scale of this imaginary, however, is equally matched by the grandness of the physical architecture which houses the main offices of the Internet Archive. Built in 1913 (Ivey, 1999, p.167), the facade of the building is constructed in a neoclassical architectural style, complete with columns, decorative iconography and three sets of iron double doors a-top a flight of steps marking the entrance. The entrance leads to a foyer that includes various stations for book and film scanning, as well as a listening station for the music contained within the Archive’s collection. The foyer leads to two

⁹<https://archive.org/details/300FunstonStSanFranciscoCa> (visited on 15th Dec. 2016)

¹⁰The Open Library is an Internet Archive initiative started in 2016 in collaboration with Aaron Swartz to ‘create one web page for every book every published’: <https://openlibrary.org/> (visited on 28th Jul. 2019)

¹¹The state of California officially designating the Internet Archive a library in 2007 (Kahle, 2007) which has wide-ranging ramifications for a US institution’s ability to claim intellectual property rights (for example, to lend and re-use digital materials) under US copyright law and fair use (Gerber, 2019).



FIGURE 4.2: The Head Office of the Internet Archive at 300 Funston Avenue, San Francisco, California.

sets of staircases that facilitate access to ‘the Great Room’, the old congregation hall, where the Archive hosts many of their large-scale events. The ceiling of the Great Room is adorned with a large stained glass dome overlooking many rows of church pews that the Archive has retained. The outer edges of the room, including ten or so rows of pews on either side of the room, are home to the artist-commissioned ceramic statues of all Internet Archive staff members who have worked at the Archive for at least three years (Figure 4.3). At a library grappling with the preservation of fleeting bits and bytes, the symbolism behind the memorialisation of a digital library/archive and the archivists who work there in stone is not lost. As Srinivasan (2016) has observed, the grandness of the setting simultaneously communicates a degree of ‘whimsy’, as well as reverence towards the Archive’s mission to capture and preserve the ephemeral.

Several of the Archive’s servers are located in the eaves of the rear wall of the Great Room, which provide a symphony of blinking blue lights that indicate when someone is accessing the Archive’s content online (Figure 4.4). Every Friday the Archive has a buffet lunch that is open to the public to come mingle with staff and volunteers, eat free food and get a tour of the Archive (often given by Brewster Kahle). Much is made of the visual spectacle of the Petabox (server) display in the Great Room during the weekly public tours of the building, which ultimately communicates the widespread utility and scale of the Archive’s servers in the circulation of information and culture online. The Petabox is the Internet Archive’s custom storage unit designed to store and process one petabyte (a million gigabytes),¹² and is representative of an overriding preference for do-it-yourself solutions that enable the non-profit to sustain

¹²<https://archive.org/web/petabox.php> (visited on 28th Jul. 2019)



FIGURE 4.3: A selection of the 100+ ceramic statues of Internet Archive staff that have worked at the Archive for at least three years. In reference to common remuneration practices within the tech industry in Silicon Valley, participants playfully referred to this tradition at the Archive as ‘statues over stock options’.

the heavy rate of digital content ingestion and access they provide on a daily basis.¹³

Moving from the ceremonial to the other working spaces in the Archive, the ground floor of the building – situated directly below the Great Room at the location of the former Sunday School – is a largely open-plan space where most of the Archive staff who work on-location are based. Senior management and directors occupy the offices that line the exterior of the open space on two sides. The main staff room contains several ‘pods’ and groups of four to six desks that are used to roughly delineate different teams and staff roles at the Archive. The physical layout of the office space, in particular, assisted in mapping and prioritising people to talk to during the course of fieldwork. For example, the Archive-It web archivists and programme managers occupied a single pod, adjacent to a group of desks where the Archive-It engineers and technical developers work, next to (but separate from) the Wayback Machine and crawl engineers. A group of couches and recliners is situated in the middle of the room which doubles as an informal space for staff breaks and a central gathering point for group meetings. The space is lined with racks and stacks of television tuners, old electronics, VHS, media readers, televisions and digital media, as well as a large television screen which enables video conferencing for the numerous staff members and volunteers that work remotely. As several informants discussed, as one of the few operational nonprofits working in tech in San Francisco, the Archive has a difficult time being able to recruit and retain staff locally given the cost of living in the

¹³In an interview at the 2016 Library Leaders Forum which I witnessed, Kahle referred to the Archive as the ‘kings of cheap’ in response to a question about their ability to cope with the expanding storage and preservation demands of collecting digital objects at scale (Ubois, 2016).

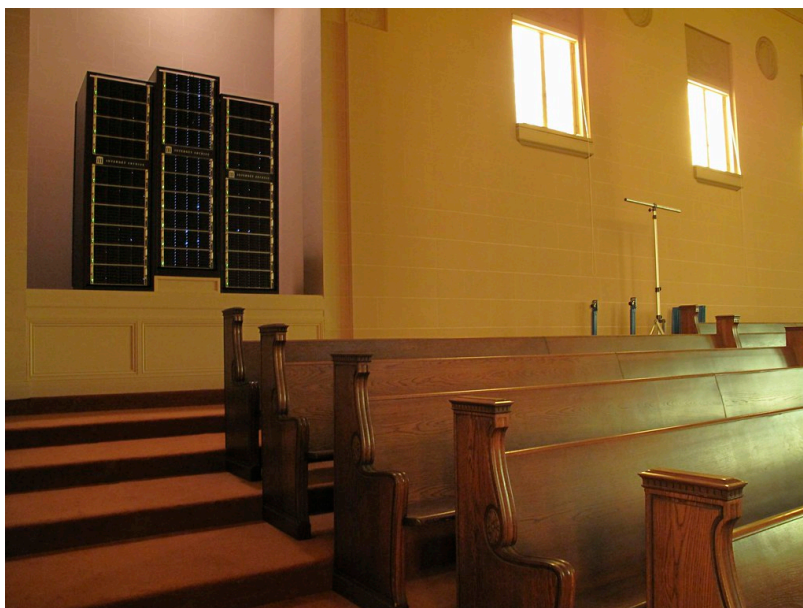


FIGURE 4.4: Petabox servers in the rear of the Great Room. Image: (Armstrong, 2011)

Bay Area. With approximately 150 employees, remote work has enabled the Archive to retain staff and volunteers who, in conjunction with those on-site, facilitate the day-to-day running and maintenance of the Internet Archive.¹⁴

4.3 Infrastructure as Labour

This section considers the ways that web archiving is shaped and sustained through the infrastructural relations of different forms of labour. The web archival activities of the Archive can be broken down into three broad areas: *crawling*, *access* and *tool development*. The Archive is engaged in crawling at many different levels and includes activities that are self-directed, those that are fully undertaken at the direction of other organisations and those crawls that are initiated and directed by other organisations using their subscription service Archive-It. The Archive facilitates access and hosting for crawls undertaken by themselves and others through the Wayback Machine, including the provision of tools that allow others to deposit and donate their own WARCs to the collection. Lastly they facilitate web archiving through their ongoing tool and technology development for crawling and replay. In practice, these three areas do not sit in isolation from one another and represent a working environment of overlapping roles, tasks and activities at the Archive. The work that makes up web archival labour permeates each of these activity areas, explored further in the following sections that consider the material ways that *knowledge work*, *translation processes* and *breakdown and repair* shape the archived Web.

¹⁴Staff figures were drawn from the Internet Archive's 2017 tax filings hosted here: <https://projects.propublica.org/nonprofits/organizations/943242767> (visited on 29th Jul. 2019)

4.3.1 Knowledge Work

Elsewhere, others have acknowledged a certain pre-occupation with abundance and ‘performing plenitude’ at so-called ‘universal archives’ such as the Internet Archive, an observation which De Kosnik (2016, p.95) argues implicitly denies, or at least distracts from, attention to any selectivity in archiving. Although it is clear that the Archive is overtly concerned with abundance and scale in their endeavours to capture more, my observations point strongly towards efforts on their part to increasingly shape, prioritise and diversify the web resources that they capture. This point was made clear at the Archive’s 20th Anniversary Party (2016) when Brewster Kahle announced that the Archive had (at that point) archived ‘273 billion webpages from over 361 million websites’ with the help of ‘Robots and 1,000 librarians’ (Figure 4.5). Here, whereas ‘robots’ refers to the use of automated ‘bots’ or web crawlers to algorithmically collect web archives, ‘1,000 librarians’ champions ongoing partnerships with university and local libraries and archives to seed collections. Kahle and other informants have identified the creation of the Archive-It subscription service as a significant step towards archiving more selectively, by providing librarians with the tools to save web resources. The Archive-It service was launched in 2006 to provide external organisations the tools to submit and curate their own web archives collections, many (but not all) of which are made available through the global Wayback Machine.¹⁵ In recent years, Kahle has regularly referenced this and previous efforts to collaborate with librarians and subjects specialists as a successful means for diversifying what gets collected in the Wayback Machine.

And yet, the selection narrative is more complex than this, or as one informant indicated when I enquired about the Archive’s appraisal practices, ‘the process is strategically mish-mashed’. Collectively, these strategies (some of which are discussed below) can be seen as one component of what Downey (2014) calls the *knowledge work*, or the ‘high value labour’ that goes into the production of information that is obscured or marketed as either automated or infinite. The notion of knowledge work is explored below as it relates to what one informant called the ‘hybrid’ crawling activities of the Archive.

One informant relayed that a common misconception about the Archive’s crawling activities is that they employ the services of ‘one giant crawler’ to archive the global Web.¹⁶ This is repeated in research literature and popular media articles around the perceived automation of the Archive’s crawling activities. At any given moment, in fact, the Archive has an (unknown) number of crawlers engaged in selective archival

¹⁵Because of the different types of content collected by institutions using the service, some Archive-It subscribers do not make their web archives collections public and are only available behind a login at the custom Archive-It Wayback Machine.

¹⁶This was actually raised in conversation around the time that I approached the Archive to do this research so should be considered ‘personal communication’.

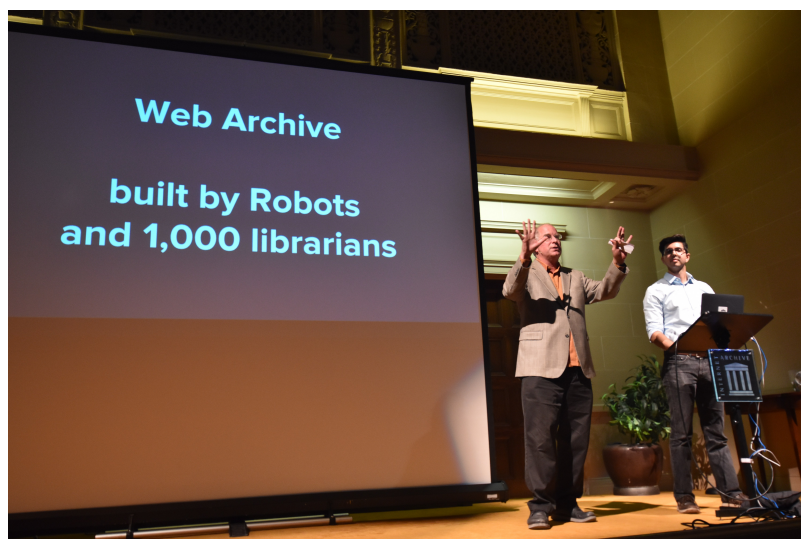


FIGURE 4.5: Brewster Kahle, founder of the Internet Archive, speaking at their 20th Anniversary Party, announces the cumulative web archival work of both 'Robots and 1,000 librarians'. Image: (Rinehart, 2016)

activities. A key priority was to begin to map these crawling events during field-work, in an effort to understand each as a contingent, sociotechnical assemblage (Table 4.1).¹⁷ These crawl events or *crawl modalities* (Summers and Punzalan, 2017) inevitably define what is collected at the Archive.¹⁸ Each is shaped by different motivations, priorities and approaches to web archiving.

Looking further back in history to 2010, a moment of significance can be observed when the Wayback team started engaging in and directing their own global crawls. Several informants described the motivations that led to this shift in direction, which was driven (at least in part) by the perceived inadequacies of the crawl data that were seeded, crawled and donated by Alexa Internet to the Archive. The Alexa crawling approach²³ is raised here as it is representative of a historical focus on the *popularity* of sites as a factor in the selection of seeds, one which (until recently) continued to influence the ways in which crawls were prioritised – even in those crawls directed

¹⁷Many of these crawl events were first established in the initial analysis of interviews (discussed in Chapter 3) as part of an effort to identify the knowledge production efforts that influence crawling activities.

¹⁸I have borrowed this term from Summers and Punzalan (2017) who describe the different ways in which crawling activities were broadly discussed and implemented in a qualitative study of web archival appraisal practices. Although different (but overlapping) crawling modalities were identified at the Archive, the focus is similarly on determining how these modalities come to pass, and how they shape what is collected.

¹⁹<https://archive.org/details/alexacrawls&tab=about> (visited on 25th Jan. 2017)

²⁰<https://archive.org/details/archiveteam&tab=about> (visited on 25th Jan. 2017)

²¹<https://archive.org/details/widecrawl&tab=about> (visited on 25th Jan. 2017)

²²https://archive.org/details/survey_crawl&tab=about (visited on 25th Jan. 2017)

²³The mechanisms behind Alexa's crawlers were presented by informants as not fully understood, and based on a historical understanding of 'how they used to work'. The proprietary nature of their crawlers and the resulting crawl data were flagged as an impediment to ever fully understanding the provenance of how resources are prioritised over others within this data.

Modality	Crawl Event	Description
Donated Crawls	Alexa CommonCrawl Cool Search ArchiveTeam	This encompasses crawls that are not harvested by the Internet Archive, but rather, donated by other organisations and community groups. Examples include the crawling activities of Alexa Internet, ¹⁹ ArchiveTeam, ²⁰ CommonCrawl, Cool Search, etc.
Global Crawls	Wide Web Crawls Survey Crawls WideOps	Starting in 2010, the Internet Archive began the Wide Crawls of the global Web in an effort to supplement the data being received from Alexa. ²¹ The Survey crawls were initiated in 2010 in an effort to take a regular snap shot of the home page of every URL in the Archive. ²²
Partner-Curated Crawls	Archive-It Crawls	Harvests initiated and curated by Archive partners through the Archive-It subscription service.
Directed Crawls	Focused Custom Contract	This is a term used by one of the informants to describe crawls that are focused on specific resource captures. Although other crawl events (e.g. Archive-It partner crawls) could fall into this category, this is specifically concerned with crawls led and initiated by Archive staff. Includes domain crawling done on behalf of contracted partners ('contract crawling').
Crowd Crawls	'Save Page Now' Wikipedia YouTube	This includes individual crawls triggered by linking events and requests, often from outside the organisation. These crawls are automated through a variety of means, including listening to the Wikipedia IRC edit channels, Twitter tweets that contain YouTube links and a number of browser-based extensions and tools that use the Save Page Now API.

TABLE 4.1: Crawl events at the Internet Archive.

by the Archive itself. The use of Alexa's 'top million' sites as the starting seeds for the 'wide crawls' (see below) was discussed as a common place practice but often resulted in the over-prioritising of popularity as an indicator of the value in capturing certain web resources – with Gregory claiming that: 'over 50% of our wide crawl was from 2,000 websites'. Further complaints were relayed about the quality of Alexa crawls as they do not capture images or embedded dynamic resources, often leading to web archives with extensive missing elements:

"[...] the quality of the captures that came from Alexa were not good enough. In the sense that Alexa was not crawling images, they were not crawling the pages that we thought we should be crawling. Because the way Alexa was crawling then—I don't know how they are doing it now—but back then was they had the Alexa toolbar and if three different users at any given day at a particular URL, they would crawl it. And that would be determined by how popular pages are, and so there was [Search Engine Optimisation] gaming happening." (Gregory, Engineer)

Various crawl events have subsequently become associated with the Archive's global crawling efforts, including for example, the 'survey crawls' and 'wide crawls'. Survey crawls are being used to supplement wide crawls by taking a snap shot of the home page of every domain/host ever identified by the Archive. Wide crawls are run twice a year over four to six months, though as Arthur described they had originally envisaged the crawl cycle to run four to six times a year. Wide crawls start from a seed list (initially the Alexa top million, as described above) and are then allowed to run autonomously with the bot following each outlink until 'it doesn't produce any interesting data any more'. When asked how wide crawls were stopped, Arthur said they have to regularly check on each Heritrix instance by manually go into the machine in question and looking at the logs to see what is actually being captured, a process they described as 'daunting'. Alex described what they were looking for when they examine the logs, which largely involves a visual inspection of the domain names contained within the capture URLs, watching out for strings that resemble 'calendar traps', pornography and endless Facebook sites. In direct result of some of the manual labour required to shape and monitor the large-scale crawls at the Archive, engineers began developing various ad-hoc tools to mitigate the need for interacting with the harvester logs and other shell scripts. One such tool is something Arthur calls the 'Domain Browser tool' used in conjunction with *Hericrawler*, a crawl queue management system the Archive developed for orchestrating large-scale crawls:

"The domain browser manual tool is for identifying undesirable domains. It's used to establish and prioritise 'shades of gray', for example only crawl this site if there are no other sites to crawl. It's used as a ranking mechanism for prioritising domains based on time, resources and place in the queue, as certain important URLs can get blocked by many instances of unimportant URLs. For example people linking to Facebook pictures can create an infinite

loop of queued Facebook links because of the nature of the graph. These types of sites are really slow to crawl as they are hosted on a single site, which must be crawled in succession because of the nature of the Heritrix crawler. Each domain/host is assigned a budget and the crawler is paused if it reaches its budget.” (Arthur, Engineer)

The Domain Browser tool is thus used to (manually) curate undesired domains based on a visual inspection of a gallery of home page thumbnails of each domain/host. The tool is set up to facilitate users tagging the site as pornography, a domain squatter or ‘link farm’ in order to remove it from subsequent crawls. Alex describes the process:

“What we did was hired half a dozen people - they would just go through it and get the top 30,000 hosts [...] and they go through 4-5,000, that’s what one person can do in a month or so. And then we get actual human interaction to say yes, this is a good website. And then we would delete or modify or prioritise based on that input. So having humans actually spend a little bit of time at the top really helped. We’d love to do it further of course.” (Alex, Engineer)

In addition, Arthur described another similar manual tool called ‘Live Update’ that they use to curate new domains that are discovered through their Wordpress crawls (crawls that are triggered by edits to sites hosted on Wordpress.com). Different to the Domain Browser tool, the Live Update tool dynamically displays the domain/host thumbnail of new domains in realtime allowing users to choose between overlay buttons tagged ‘P’ for pornography, and ‘F’ for link farms, or visit the site for further investigation. Arthur said it was developed in an effort to ‘gamify’ the process of curation and described using the tool whenever he had downtime. When asked how to spot a link farm, Arthur responded that ‘it’s obvious, there’s usually a giant box [iframe] with keywords and a list of domains on the home page, easy to spot’. The use of manual curation tools reveals both the role of human intervention in the process of curating millions of links and the tensions and trade-offs that exist between the use of bots and a desire to capture ‘high quality’ sites.

In response to the restrictive number of URLs collected by the wide crawls seeded with the Alexa ‘top million’ sites, several informants described some of the Archive’s more recent efforts to study the wide crawls through a grant they received to improve the Wayback Machine in various respects. One such study of ‘wide crawl 12’ is captured in a grant milestone report (that was made available) which describes the various techniques used and makes recommendations for improving future crawls (Goel, 2016b). Here, Gregory describes the process of studying the hyperlink structure of existing archives to seed crawls:

“I do a lot of link analysis where I study the hyperlink structure of our crawls and try to figure out in certain pockets, use some rank methodologies to figure out ‘oh these are important resources’, for instance they have a lot of

links to them or traffic is really high – let’s seed the crawl with those. The most recent wide crawl I took the most linked to pages from every single website, so 230 million websites [...] and instead of crawling the Alexa top million, let’s crawl this bit. Sort of like a hybrid survey and wide model [...]. And we found resources that we had never crawled before. I’m not saying one is better than the other; I’m saying that hybridising this process might be one way of balancing the scales a little bit.” (Gregory, Engineer)

This type of link analysis thus assists in finding the edge nodes of websites that the crawlers have identified – sometimes upwards of 60 million sites – but do not get around to crawling before the crawl is stopped. These are then used as seeds for the survey crawl (to crawl the home pages of the sites) and to iteratively expand the net and number of websites captured by the Archive. Gregory indicates that through these types of studies, they estimate that at any given time they are only crawling around 20% of the Web. Gregory structures the issues surrounding balance in selection priorities as a problem of resources (a theme which repeatedly arises), but outlines three considerations for determining ‘better crawls’ and ensuring they are crawling the ‘right 20%’:

*“The way I think of it [...] there’s three branches, there’s **popularity**, there’s **novelty** and there’s the **risk of going away**. How do you achieve that balance? You want to get stuff that people are using - not just junk that is on the Web that you’re just filling up the servers with that won’t ever be found useful, like calendar pages, things like that, crap [...] there’s no novelty. It’s new? We want to make sure it’s preserved because it just came out, it’s a new article, it’s a new website. And then there’s the risk of going away [...] – if you’re going to shut down this service—Vine is going away—we jump in and crawl. So as we’re crawling the Web can we do a good job of sort of achieving that balance? We don’t quite know what the solution is to achieving that balance.” (Gregory, Engineer, emphasis my own)*

If we expand the picture to look at some of the other crawl modalities of the Archive, the multi-faceted approach to selection becomes even more apparent. A number of techniques for selecting domain/hosts were described by participants associated with the Archive’s contract crawling, or the custom and domain-level crawling undertaken on behalf of partner institutions. A manager, Elaine described the use of zone files, partner-submitted seed lists (via Google spreadsheets or forms), links embedded in particular social media streams, and using geographical look-ups of existing content held by the Archive to extract relevant domains for crawling. Other sources for selection include ‘listening in’ on Twitter to determine which YouTube videos get linked to, as well as what outlinks get added to Wikipedia – both of which trigger crawl events. Increasingly, the Archive has also been developing a variety of tools that use their longstanding ‘Save Page Now’ feature to promote the saving of web resources to

the Wayback Machine by anyone with access to the archive.org home page, a Firefox plugin or the API.

These methods highlight some of the ways the Archive is leveraging the power and labour of ‘the crowds’ – through the users of Twitter, Wikipedia and Save Page Now – to not only diversify and ‘balance’ the domains and types of resources that are archived, but also (implicitly) co-opt and transform these users into potential stakeholders of the Archive. Furthermore, the Archive is in multiple ways, leveraging the web archives amassed in the Wayback Machine project over the last 20 years to continue to increase their net resources. Here, web engineers use the Survey crawls to find and re-use links that are saved by other crawling methods (like the crowdsourced ones described above) to seed and expand subsequent crawls. This practice could be seen as a form of *knowing capitalism* (Thrift, 2005), where the networked labour involved in amassing web archives at scale becomes an ‘infrastructural logic’ that works to generate further capital in the form of a bigger and bigger archive.

The next section explores the ways that upon selection, these web archives are transported and transformed at the Archive.

4.3.2 Translation

Several informants detailed the various processes that enable the transfer and preservation of live Web resources into archived resources at the Internet Archive. Drawing again on Downey (2014), I want to consider how aspects of web archival labour act as a form of *translation*, where practices work to not only move web resources from the live Web to archives, but also invariably transform the meaning and nature of resources themselves. Using several examples of ‘real time stenography’, or the processes associated with transcribing ‘ephemeral human speech’, Downey (2014, p.155) describes how stenography involves translation processes that implicate stenographers in making the ephemeral accessible across different spatial and temporal contexts:

“[...] [these] are examples of what we might call *translation processes* in a broad sense, because in each case they involve not just moving information unaltered from one set of technical codes to another (like in telegraphy), nor just creating the contextual environment for information to circulate from one institutional or intellectual context to another (like in librarianship), but a sort of recasting of the very meaning of the information content in the first place” (Downey, 2014, p.158).

Here, Downey is referring to the interpretation, negotiation and ‘editorial interventions’ that enable the translation of human speech into fixed and permanent representations of the ephemeral. They argue that these practices require the creation of an ‘environment’ or infrastructure to enable the flow and interaction of information

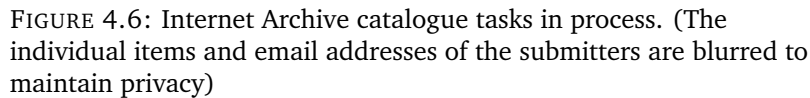
that as a result, often ‘recasts’ the meaning of targeted resources in the process. Here, examining web archiving labour as not merely transfer processes that copy HTML content from the live Web to the Archive’s servers – but rather as *translation* processes – places emphasis on the contingent and constructed nature of web archives, and materially connects the archivist and the infrastructure (labour, tools and technologies) of archiving to the ways that web archives can be understood.²⁴

The Wayback Machine itself provides a starting point for examining some of these translation processes and their implications for the archived Web. The underlying infrastructure that facilitates the Wayback Machine was described as a complex series of interconnected processes that both ingest and make web archives available online – succinctly summarised by one informant as: ‘input, process, output’. Several translation processes were identified during analysis that occur to digital objects in the Archive (e.g. compress, transfer, decrypt, derive, de-duplicate, exclude), some of which can be seen as individual rows in a screenshot of an example queue of tasks in the administrator’s view of the Archive’s catalogue system (Figure 4.6). In order to facilitate replay and access, several processes are performed on web archives that are dependent on the source and collector of the WARC files being ingested. For example, a decryption process is performed on Alexa Internet donations to enable access to embargoed data (after six months) within Wayback. All files that enter the Archive go through what is called a *derive* process that is dependent on the filetype entering the network. For web archives, this enables the automated indexing of WARC files to produce the ‘layers of indices’ that enable resources to be located and accessed through the public Wayback Machine:

“We have this massively distributed file system that has 20,000 hard drives on 500 plus nodes that are all part of the fabric that collectively represent a whole lot of storage that is highly available and highly accessible. [...] In the case of the WayBack Machine [...] we get these WARC files and we know they are from websites, we turn them into index files. [...] The WayBack Machine works based upon a whole lot of rules and are based on this set of four layers of indexes that provide pointers to actual web captures that are sitting on hard drives. So when I go and try to playback a page, the actual assets that are going to represent that page – the individual web captures may and often times do live on different physical machines. Maybe even different locations too.” (Nicholas, Engineer)

Nicholas describes a complex system of physical storage, layers of index files and a ‘whole lot of rules’ that underlie the Wayback Machine. This description supports and neatly illustrates the relational qualities of web archiving as infrastructure. It highlights the integral and performative role of the Wayback Machine in the active

²⁴This reinforces previous points about the contingencies inherent in the construction of web archives, but also the ontological differences inherent to web archives as ‘re-born digital’ artefacts (Brügger, 2012, 2016).



The layers of indices and associated rules used to query the archives are worthy of some further attention as they also form another aspect of the process of translation, namely the ways that social protocols are embedded and enacted through technical protocols of access. The exclusion rules that govern access to web archives stored in the Wayback Machine have received increased attention in recent years through public discussions of the continued utility of the `robots.txt` protocol (Koster, 1993) and its role in enabling and preventing access to the archived Web at the Internet Archive (Scott, 2017; Summers, 2017). The `robots.txt` protocol provides website owners with a mechanism to communicate both whether or not Archive crawlers should be allowed to crawl web content, but also whether or not the Archive should make that content available through the Wayback Machine. Again, Nicholas explains:

*of them are accessing their own sets of exclusion tables because as part of what we do, we don't just take, capture and present. We capture, process, present. Part of the processing may be that we have received a DMCA request – Digital Millennium Copyright Request – to say that 'I'm the owner of this website, I don't want it to be available via the WayBack Machine'. We honour that. Or there's a robots exclusion – so someone is saying disallow * useragent [ia_archiver]. (Nicholas, Engineer)*

Here as Nicholas describes, the Wayback Machine is subject to DMCA requests and robots.txt exclusions that at least, in theory, determine the ways that they provide access to certain archived sites and pages. To some extent, the legal and ethical protocols governing the archiving of the public Web have always been in question, as exhibited by Kahle's description of the mood surrounding their initial attempts to archive the Web back in 1996, remembering that 'all our lawyer friends said that if we collected the web, lawyers would rain down on us like frogs'. A milestone came in 2002 when, only months after the initial unveiling of the Wayback Machine, the Internet Archive became one of several web publishers to be implicated in wide-ranging attempts by the Church of Scientology and their lawyers to remove access to web content that was seen to be critical of the church (Bowman, 2002; Jeff, 2002). Subsequently, and in consultation with other free speech advocacy organisations such as the Electronic Freedom Foundation (EFF), the Archive sponsored the construction of the Oakland Archive Policy (OAP) that produced a series of recommendations for operationalising situational responses to requests for the removal of online access to archived web content.²⁵

Despite the OAP recommendations however, the ways in which the Internet Archive makes decisions about who and what to provide access to remains somewhat shrouded in mystery.²⁶ Decisions surrounding the take-down refusals in the case of the Joy-Ann Reid blogs (discussed in Chapter 1) coupled with recent announcements that the Archive would no longer obey robots.txt exclusions on US government websites (Graham, 2017; Rossi, 2017), have demonstrated the ways that the Wayback Machine's sociotechnical protocols of access are relational, only temporarily stable and always subject to change. Here, the material relations of infrastructure (as people, practice, software, hardware) act as arbitrators in the performance of the politics of accessing the Wayback Machine. These relational entities are inevitably translated through the situated technical, legal and ethical frameworks within which they are based, and ultimately work to construct both how the Web is archived and indeed how access to the Web's past is enabled.

²⁵<http://groups.ischool.berkeley.edu/archive/aps/removal-policy> (visited on 29th Jul. 2019)

²⁶On several occasions when I pressed discussions about the robots.txt protocol and the ways that decisions are made surrounding access to archived pages I was reminded of the constraints of staff non-disclosure agreements and referred to the Wayback Machine's FAQs: <https://help.archive.org/hc/en-us/articles/360004716091-Wayback-Machine-General-Information> (visited on 29th Jul. 2019)

4.3.3 Breakdown, Maintenance and Repair

Many informants described scenarios where human mediation was required in otherwise (seemingly) automated processes. Maintaining the technical infrastructure of web archiving may be akin to Coleman's description of the role of system administrators as 'part plumber, part groundskeeper, and part ninja, fixing problems, maintaining the system, and fending off attacks' (2012, p.12). Here, informants described the daily tasks of monitoring and repairing the network of crawler nodes and 20,000 hard drives (Gonzalez, 2016) – to replace failed disks, tripped power cords and mitigate heavy network traffic (Figure 4.7). Other examples require the manual intervention of engineers and support staff focused on crawling, including: running *patch crawls* when bots failed to crawl designated seeds, or in the event of a *crawler trap* where bots are trapped in an infinite loop of seed requests, or when *missing* or *altered* elements are observed in the playback of archived web pages. Fundamentally, each of these boil down to issues surrounding either the capture or replay of web resources. Borrowing from Star and Ruhleder (1996), these moments could be considered a form of *breakdown* which reveal the infrastructure beneath, the contingent assemblage of processes that – for example, enable the capture of intended web resources or provide 'high fidelity' access to the archived collections.

As discussed in Chapter 2, issues surrounding the *quality assurance* of capture and playback are not unknown dimensions of web archival practice, however, the processes that are undergone to mitigate these issues are not well documented. Here I draw on the work of Jackson (2014) who advocates for an examination of the moments of *breakdown*, *repair* and *maintenance* of technologies, in that they redirect attention to the act and 'ethics of care' and embody the creation of value by their maintainers. In other words, the practice and processes involved in fixing and maintaining technologies can be used as evidence for their worth by those who sustain these practices over time. These moments of maintenance and repair in web archiving are present throughout the study, including the repair and maintenance of crawl data and crawling technologies, Wayback and access tools, as well as the repair of broken links on the live Web by the Archive. Some of these practices are discussed further below, as they highlight some of the factors that influence the decision-making and technical processes that enable the repair of web archives (and the technologies that enable access).

Repairing Web Archives

A few training sessions for a web archivist on the Archive-It team, Karen were observed. Through listening to Karen's Q&A with other team members, certain junctures were highlighted where support staff are regularly required to prioritise activities, particularly in response to the 'quality' or 'completeness' of web archives (and

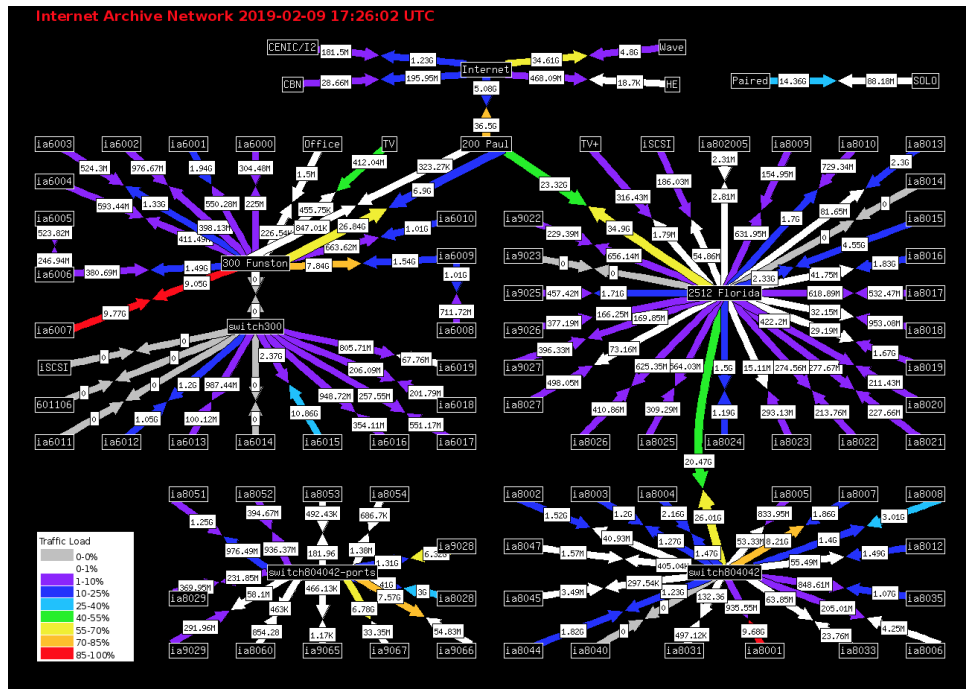


FIGURE 4.7: The Internet Archive Weather Map shows network traffic in and out of each component of the server infrastructure.

<https://monitor.archive.org/weathermap/weathermap.html>
(visited on 9th Feb. 2019)

as raised by partner organisations). To set the scene a bit, I have drawn on my interpretive notes following an observation session:

In the second session we all sat on the couch in the pit. Lydia asked Karen if she wanted to continue the training they began in the morning, which was aimed at addressing a recent support request that came in from a partner. When the conversation got a bit technical – as a result of Karen asking increasingly detailed questions about how certain seeds and test crawls are rendered in playback – Mike (a support engineer) was called over by Lydia. Mike walked over and leaned on the couch and began to explain some common differences between capture and playback issues. The consensus from Lydia and Mike seems to be that the first task in the support role is to determine whether the issue is related to capture or playback.

Karen is concerned about waiting to solve issues based on partner requests, advocating for the team to be more actively QA'ing collections in case they are at risk of disappearing. Mike responds that there are different issues at play (including time and resources) and that it's key to understand that capture issues will always take precedent over playback issues – for exactly that reason.

From this vignette I observed (and interpreted) several points that led to further questioning in subsequent interviews. First, we can see that Karen is getting to grips with

a key aspect of the role of the web archivist at Archive-It, which (in conjunction with support engineers) is to determine the difference between playback and collection issues and respond accordingly. Second, the comment that capture problems will always trump replay problems is insightful. It emphasises the goal of capture and reflects the underlying motivation driving activities – the fear of disappearance. This observation also emphasises the active role of the web archivist and support engineer in the processes that shape the ‘fidelity’ of web archives. They are implicitly driven by time and resource constraints (a fact that was repeatedly brought up by informants) but they are also active participants in the practice of choosing which support issues are prioritised.

Around this question of prioritisation, in an interview Mike reflected on a number of factors that influence how web archiving support tasks are prioritised in practice. Mike first explained that Archive-It employs ‘agile development practices’²⁷ and that they use a ticketing system to log and keep track of both internal and external support requests:

Mike: *“Going back to our agile discussion, in our daily ‘stand-up’ we’re picking from what we do in this sprint (which is a two week sprint). We have a whole list of items that are either high priority because the content that we (or partners) are looking to capture are high priority content or it’s a high priority partner for one reason or another.”*

Jessica: *“What reasons would the content be high priority?”*

Mike: *“Something in jeopardy of going away on the Web is one really good example. Or in the case of whitehouse.gov [...] yes, it’s going away but regimes are changing, the site isn’t disappearing but we know a lot of the content is going to change so we’ll go out and capture that as quickly as possible. It’s [Joan]’s job and a lot of the web archivists to go and put things in prioritised backlog and Joan and others will pull from that to form our sprint [...]”*

Mike indicated that specific repair tasks are prioritised in each daily ‘standup’ where team members report on upcoming development goals. Web archivists and programme managers will mark tasks as high priority either because the content is at risk of changing or going away, or because the request comes from a high priority partner (as both were the case for the whitehouse.gov and End of Term archives that were being captured at the time).

²⁷Agile software development comes in many flavours, but in this context the informant is referring to a common practice called a ‘stand-up’ which involves a short ‘time-boxed’ meeting session where team members report on their progress and impediments to the upcoming code development goals (Yip, 2016).

Repairing the Web

Further examples of repair can be found in the work that the Archive has been engaged with in fixing broken links contained in Wikipedia as part of their ‘No More 404s’ Project. Through the use of automated bots and manual intervention, the Archive (and collaborators) have been altering URLs (out-links) on Wikipedia with a 404 status code to point towards archived versions within the Wayback Machine. Nicholas described the Archive’s role in this process:

“[...] [We] look at every link in English WikiPedia and we’ll check the status code. So if it’s not a status code 200 or maybe if it’s a status code 404 [we] say – can we edit that link to point to a capture of that web page available via the WayBack machine back when it was a 200. So that’s what we did. And we did it – when we say ‘we did it’ – two volunteers did it actually. It was amazing. [...] And they wrote a robot to go crawl through Wikipedia and basically check the status codes and look those pages up in the WayBack Machine and attempt to do the right thing. [...] what we said recently is that we repaired more than a million links. The nuance of that is that 600k of those had to actually be repaired by human beings over the course of many years using the Wayback Machine. Like 15 years or whatever. The last 400k links we helped repair over the last few months.” (Nicholas, Engineer)

On their own, these activities deserve further investigation to document the motivations and practices which were undertaken to perform these changes. As it relates to this research, the Wikipedia changes demonstrate how the Archive is engaged in not only archiving at scale, but also actively changing the landscape of the live Web by making itself one of the primary mechanisms for accessing the historic Web. However, these types of changes to Wikipedia source references are not without controversy. There is a perception that this maintenance work has further entrenched a reliance on the Archive as the sole provider of archived web pages within the landscape of the Web. Some have advocated for a more ‘robust’ approach to link repair, for example, using the Memento aggregation protocol to check for archived resources in multiple sources rather than just the Archive (Rosenthal, 2016). This observation is bolstered by the number of controversies surrounding the use of bots to automate changes to Wikipedia 404s to point towards another web archive, Archive.is²⁸ which resulted in the banning of a number of bots and users and four lengthy Wikipedia Requests for Comments (RFCs).²⁹ The wholesale and continuing practice of replacing Wikipedia outgoing (404) links provides further evidence of an increasing reach and reliance on the Archive as part of the Web’s architecture. The issues surrounding link repair fundamentally raises further questions about the role of power in the maintenance of

²⁸<http://archive.is>; also located at <http://archive.today>

²⁹https://en.wikipedia.org/wiki/Wikipedia:Requests_for_comment/Archive.is_RFC_4#Previous_RFCs (visited on 5th Feb. 2017)

archived resources, and the situated decision making processes that determine which versions of the Web are to be the canonical versions of the past.

4.4 Chapter Summary

The findings of this chapter offer some important insights into the research question: ***In what ways do web archival practices (the who, why and how) shape the archived Web?*** *Web archiving as infrastructure* enabled two interconnected and overarching observations about the ways that web archiving shapes the archived Web. The first is that web archiving is composed of a dynamic set of spatially and temporally located practices that shape the nature of what and how the Web is archived. This observation was demonstrated through my efforts to locate and contextualise the emergence of the Internet Archive within the 1990s and early 2000s in Silicon Valley. Here, I discussed the early goals of the Archive and its founder Brewster Kahle to build the online ‘Library of Alexandria 2.0’, in pursuit of particular sociotechnical imaginaries to enable ‘universal access to all knowledge’. I extended the view of this digital library to a description of the physical space in which the Archive occupies, in an effort to further situate the place in which local web archiving practices are shaped. In the round, the first part of this chapter worked to situate the ways that web archiving as infrastructure, and the practices this enables, emerges within particular social, cultural, economic, legal and ethical contexts that shape the material ways that web archiving is approached and enacted.

This observation is further supported by the analysis of the components of web archival labour through the second half of this chapter. Extending Downey’s (2014) concept of information labour, I outlined the case for four components of labour (knowledge work, translation, maintenance and repair) that support the relational view of web archiving as infrastructure and emphasise the explicit ways that the everyday work of web archivists and engineers at the Internet Archive are shaping the archived Web. This analysis points towards a complex system of knowledge and maintenance work for prioritising which web assets to collect and repair. The Archive is leveraging their extensive existing archives for understanding networked linking behaviour in an effort to balance the breadth and depth of crawling activities, while discovering new sources for identifying websites to crawl based on measures of popularity, ‘novelty’ and sites that are endanger of going offline. The team has devised multiple mechanisms for identifying different types of ‘undesirable domains’, including rule-based link pattern-matching and the development of ‘gamified’ tools for the manual curation of sites.

Collectively, the efforts of the Archive can be seen as knowledge work, and these activities, seen in combination with other practices around the prioritisation, repair and maintenance of tools and archives all have ramifications for how web resources are

transformed for use. It is the labour of non/human agents that enables the preservation and ingestion of information from the Web into the Archive, and then once again back to the Web where archives are reassembled via the Wayback Machine. Although imperfect, this labour is increasingly recognised as an essential element of the web architecture. The information labour and knowledge work of potential web archival users is therefore intimately tied to the web archival labour of the Internet Archive. As the global Wayback Machine currently provides access to billions of webpages – often inaccessible elsewhere – editorial decisions have implications for not only the fidelity of archived captures, but indeed whether or not certain parts of the Web are preserved at all.

The next chapter, Chapter 5, expands the field of view to consider *web archiving as culture* through the case of the Archive Team, a community of ‘rogue’ archivists.

“Culture, then, consists of standards for deciding what is, standards for deciding what can be, standards for deciding how one feels about it, standards for deciding how to go about doing it.”

WARD H. GOODENOUGH (1963, *Cooperation in Change*)

“I did it because I had the means (disk space), the motive (the sense of history and the recognition that this was historically relevant work representing thousands of hours) and the opportunity (a fast connection and five days before they were to die).”

JASON SCOTT (2009, *Datapocalypse!*)

5

Web Archiving as Culture: Archive Team

5.1 Introduction

This chapter explores *web archiving as culture* through the case of the Archive Team, ‘a loose collective of rogue archivists, programmers, writers and loudmouths dedicated to saving our digital heritage’.¹ Here, *culture* is used as a vehicle to explore how web archives are produced and enacted through systems of cultural practices. The concept of culture helps to delineate two distinct, yet interlinked observations about web archiving. The first is that web archives, like culture, are produced and enacted through practice. Here, the example of Archive Team and their efforts to archive Tumblr are used to demonstrate that web archiving is situated within particular cultural worlds that are observable through local practices and associated systems of meanings. The second observation positions web archiving as a transformative force that also produces culture. Here, I argue that web archiving practices both create the archived Web and reinforce the cultural worlds and communities they inhabit. In the case of Archive Team therefore, participation transforms web archives through practice and produces a culture of affinity around the quest to archive the Web.

Culture as Practice/Practice as Culture

The first observation is rooted in a view that culture is the articulation of practice and systems of symbols and meanings (Sewell, 1999), or in other words, culture is

¹<https://www.archive-team.org/> (visited 26th Jul. 2019)

conceived as something that people do. I draw parallels and inspiration from Seaver (2017) who positions *algorithms as culture* in an effort to foreground the ways that conceptually, algorithms are ‘multiples’ (Mol, 2002), or unstable objects that are enacted through varying forms of cultural practices. Rather than solely favouring a technical, or engineering-based conceptualisation of algorithms, Seaver emphasises the numerous and everyday ways that algorithms are produced through the ‘local production of abstract representations’, ‘human sensemaking’ and a diverse and changing set of sociocultural contexts (2017, p.6). Similarly, and in line with a practice theory approach to culture (Bourdieu, 1977; Ortner, 1984, 2006), I argue that web archives are constituted through ‘strategies of action’ (Swidler, 1986) which are themselves articulations of meaning-making through practice. Culture therefore, as a diverse *system of cultural practices*, invokes a ‘toolkit of symbols, stories, rituals and world-views’ (Swidler, 1986, p.273) that shapes how the Web is archived. In this case, the practices or strategies of action employed by Archive Team are informed by particular tenets of practice that frame an importance in the abundance of collection, distributed working practices and a general reliance on irony and humour to mobilise participants. Articulating these practices and the ‘toolkit’ that informs them is necessary for considering the impact of practices on the nature of what is archived.

The second observation is that web archiving is transformative; where a focus on cultural practices works to frame the ways the Web is altered through archival practices, as well as highlights the contribution of culture ‘in sustaining existing strategies of action and its role in constructing new ones’ (Swidler, 1986, p.278). As discussed earlier in this thesis, we know that web archiving alters the nature of what is archived through the process of archiving (Brügger, 2012, p.108), and in this chapter I extend this observation to include the ways that culture is implicated in this process of transformation. I argue that culture works to define, shape and transform the nature of what is done, as well as create communities around the doing of web archiving. Here, web archiving (and the Web itself) works to bring together and sustain communities of affinity through the generation of code, stories, symbols, practice ‘rituals’ and ways of viewing the world that reinforces their mission of archiving the Web whilst also materially shaping how the Web is archived.

Introducing Archive Team

Although featured in numerous blog posts, popular media articles, public presentations, podcasts and soundbites since 2009, the specifics regarding the processes that underpin Archive Team activities have yet to receive critical attention within the wider web archival domain. The Archive Team is distinct from other web archiving initiatives in several respects which will be discussed, including the implications of being based outside of the professional, institutional and legislative constraints of memory institutions typically represented in web archiving research communications.

Archive Team and its participants have developed bespoke archival pipelines that take advantage of distributed computing resources and volunteer labour to archive sites and platforms at scale. Through the use of IRC logs, documentary sources and interviews, this chapter reveals some of the ways in which Archive Team simultaneously works *with* the standards of conventional web archivists to extend these technologies to suit their own needs; whilst also *rejecting* risk-averse, professional (library/archival) norms that tend to dominate discussions of web archiving practices within the field.

This chapter is presented in two halves that work to frame *web archiving as culture*. In the first half I use the notion of community to frame the cultural dimension of Archive Team's web archiving practices. Here I argue that practice is structured through the ways that Archive Team materially and symbolically constructs a community of practice. I then introduce Archive Team through a discussion of their origin story and development of community protocols that work to enable web archiving as a form of collective action. The chapter discusses how IRC and the wiki work to organise, mobilise and enrol new participants into the work of archiving the Web. I then use the analogy of a radical environmental activist group, EarthFirst! to outline and propose two tenets of 'radical web archiving' that shape the ways in which Archive Team frames web archiving demands an approach that champions a form of what I call *archival neutrality* and *brute force archiving*.

The second half of this chapter presents an account of Archive Team's attempts to archive Tumblr NSFW ('Not Safe For Work') blogs between December 2018 - January 2019.² I followed the activities as they occurred on public IRC channels (#archiveteam, #archiveteam-bs and #tumbledown), Archive Team's GitHub repositories, social media (Tumblr, Twitter, Reddit) and in coverage by popular news media sources. Further context is added to observations which draws on historical IRC logs, previous projects undertaken by Archive Team and interviews. Although it is clear from this research that Tumblr presented a particular set of circumstances for archiving – constrained by the sociotechnical affordances of the platform – this account explores the ways in which an enterprising Archive Team (as a distributed, dynamic collective of 'rogue archivists') shapes and navigates the challenges of archiving the Web.

5.2 Framing Archive Team, Constructing Community

This section takes its inspiration from Cohen's (1985) work on the symbolic ways that people construct communities. Here, community is used as an analytical device through which to frame and probe the cultural dimension of Archive Team's web

²Not Safe For Work (NSFW) is an internet colloquialism for links and content that may contain nudity, sexuality, profanity or violence and by implication, not suitable for viewing whilst in a public place or workplace setting.

archiving practices. Community is positioned as a *relational* idea, where communities are formed (and sustained) through notions of both commonality and opposition; similarity and difference (Cohen, 1985, p.12). Particular significance is given to the construction of *meaning*, and by extension (in the context of this practice-based approach to culture), the ways that meaning is produced through web archiving practices. As such, attention is paid to the ‘webs of meaning’ (Geertz, 1973) or symbolic forms and manifestations of meaning, where symbols, stories and ‘rituals’ are used by Archive Team members to motivate and constrain participation, create belonging, as well as frame struggle and success. As Cohen puts it:

“[...through] symbolic behaviour, people draw the conventions of community about them, like a cloak around the shoulders, to protect them from the elements – other people’s ways of doing things, other cultures, other communities. The conventions become boundary through their re-investment with symbolic value” (Cohen, 1985, p.63).

As a loose ‘anti-establishment’ collective of volunteers, I argue that in fact, Archive Team creates organisational norms and shared notions of membership that drive web archiving through the use of irony, humour and metaphor, distributed working strategies and the transmission and enforcement of their own practice conventions. Many of Archive Team’s ‘conventions’ undoubtedly serve practical purposes for an online community centred on virtual participation, for example, in using distributed web technologies for organising and managing remote communication and archiving. In addition however, I propose that practice conventions are also symbolic of a broader attentiveness to sociotechnical values that can be associated with Levy’s (now famous) definition of the ‘hacker ethic’ (Levy, 2010). Based on their observations of computer programmers at MIT in the 1950s and 60s, Levy outlines a ‘hacker culture’ and set of core tenets that emphasises: a general commitment to information freedom and freedom of expression, anti-bureaucracy and a mistrust of authority, the meritocracy of hacking and the aesthetic potential of computers for making a better world (2010, pp.39-46). Similar to the F/OSS software programmers in Coleman’s (2013) study, Archive Team should also be seen as a loose ‘composite of distinct yet connected moral genres’ that simultaneously critiques and aligns themselves with the broader tenets of liberalism that often pervades hacker communities (2013, p.19).

Below, the origin story of Archive Team is first used to contextualise the early visions for a volunteer ‘emergency response team’ dedicated to archiving online user-generated content. The origin story, accompanying analogies and imagery are used to frame both ‘the opposition’ (corporations, gatekeepers, censorship) and an ‘A-Team’ imaginary with a moral cause to save the Web’s history. Next, I frame the conventions employed by Archive Team to organise and enact web archiving, as well as to transmit practice protocols to new participants. As will be discussed further through the work of archiving Tumblr, the ways these practices and strategies are enacted, but



FIGURE 5.1: Archive Team, a rogue band of activist web archivists formed in response to the shuttering of online hosting services for user-generated content. (Image: Jason Scott, <https://www.archiveteam.org/index.php?title=File:Archiveteam.jpg> (visited on 27th Jan. 2019))

also contended and negotiated through action, ultimately work to shape the nature of how the Web is archived.

5.2.1 An 'Emergency Response Team' for Web Archiving

Co-founded by Jason Scott Sadofsky,³ more commonly known as Jason Scott, Archive Team describes themselves as 'a loose collective of rogue archivists, programmers, writers and loudmouths dedicated to saving our digital heritage'.⁴ The collective formed in January 2009 in response to a series of blog posts by Jason Scott following the closure of AOL *Hometown*, an early web-hosting platform which allowed AOL users to build websites with little-to-no coding expertise (Hogan, 1995). The closure and loss of AOL Hometown in 2008, a service that hosted 14 million websites as of 2002 (Hu, 2002), was likened to 'a mass eviction' (Scott, 2008) that amounted to a loss of people's 'information, hopes, dreams [and] history' (Scott, 2009a). The closure became a rallying call that subsequently spawned numerous web archiving projects for Archive Team, including partial captures of GeoCities, Friendster, and Google Video, to name only a few.⁵ Jason Scott's initial vision for the collective is illustrated in this blog post excerpt:

"ARCHIVE TEAM would be like CERT (the Computer Emergency Response Team) used to be, where it was a bunch of disparate people working

³Others, like Joey Hess have also indicated that they were founding members of Archive Team (Hess, 2015).

⁴<http://www.archiveteam.org> (visited on 9th Apr. 2019)

⁵<http://www.archiveteam.org/index.php?title=Projects> (visited on 10th Jan. 2017)

together to solve a problem in a nimble and networked fashion. They'd find out a site was going down, and they'd get to work.

They'd go to a site, spider the living crap out of it, reverse engineer what they could, and then put it all up on archive.org or another hosting location, so people could grab things they needed. Fuck the [End User Licence Agreements] and the clickthroughs. This is history, you bastards. We're coming in, a team of multiples, and we will utilize Tor and scripting and all manner of chicanery and we will dupe the hell out of your dying, destroyed, losing-the-big-battle website and save it for the people who were dumb enough to think you'd last. Or the people who, finding you'd been around forever, had the utter gall to *not be near their computers* during your self-created, arbitrary sunset period" (Scott, 2009a, emphasis in the original).

This excerpt from Scott's original vision provides a working sociotechnical imaginary for Archive Team, as well as symbolic commentary on the loss of web history and the erosion of the rights of platform users. Scott draws a direct comparison to CERT, a group of computer experts at Carnegie Mellon University that was established in 1980s to convene and respond to computer security incidents. Here, Archive Team is envisioned to be an 'emergency response team' that dedicate their skills to archiving websites *for the users*; an 'A-Team' in the battle against corporations who do not provide users with the means to save their own content for posterity.⁶ The vision prioritises a distributed, 'nimble' and 'networked' team and foregrounds a sense of loss ('this is history') and deep mis-trust in hosted web services. Scott's vision for Archive Team continues on, and paints a vision of 'vigilante teams of mad archivists' for the Web, who work to not only archive digital content, but also to 'publicize [the] demise' of failing platforms (Scott, 2009a). Motivated by a desire to preserve web-based ephemera and what they frame as 'our digital culture' for the future, Archive Team exhibits a strikingly hacker-activist orientation to the use of web archiving as a means for advocating for the rights of platform users. This point is illustrated in another quote from Jason Scott referring to the early aims of Archive Team:

"The goals are myriad but I think the easiest one to achieve is to highlight and [embarrass] companies that take a cavalier attitude to removing user data with extraordinarily short notice" (Scott, 2009b).

In the past, Jason Scott has referenced the GeoCities project as the project that 'made' Archive Team as it gave Archive Team publicity opportunities and mobilised people around archiving the Web (Findlay, 2011). Scott argues that their 'overly aggressive' approach and use of humour worked to make people enthusiastic about getting involved in the project. This approach can be observed through efforts to publicly 'name and shame' companies and platforms (in the media and on their wiki) that

⁶*The A-Team* was an American 1980s television series that featured a group of mercenary, 'soldiers of fortune' hired to solve problems whilst being on the run from the military police.

resist the crawling activities of Archive Team and/or refuse to give users tools and reasonable time frames for exporting their content. Web archiving is therefore positioned as a moral good, but also a form of resistance and civil disobedience that works to mobilise participants to the cause.

5.2.2 Archive Team as Community Protocols

Following the culture as practice approach, web archiving as culture is observable through the practices used to both action and organise the collective work of web archiving. Below I discuss some of the ways that community identity is conveyed through sociotechnical protocols and tools that facilitate a form of ‘institutional memory-making’ through self-archiving, shape remote communication and project organisation, create a sense of community belonging and ultimately reflect particular aspects of ‘hacker culture’ that shape web archiving in practice. The conclusion of this section surfaces and proposes two tenets of practice – what I’m calling *archival neutrality* and *brute force archiving* – that when actioned produce both community guidance and particular practice dilemmas for Archive Team, both of which will be further considered through the project of archiving Tumblr (Section 5.3).

‘Rituals’ of self-archiving

Immediately following Scott’s initial ‘call to arms’, a MediaWiki instance was installed at archiveteam.org and the *Internet Relay Chat* (IRC) channel #archiveteam was created on the IRC network, EFNet. IRC was invented in the 1980s, and as the name suggests, it is the foundational mechanism for facilitating distributed internet-based group chat, private messaging and file sharing. As noted by Coleman (2014, p.8) in their ethnographic study of *Anonymous*, IRC is often the preferred communication device for geek and hacker communities online.⁷ In keeping with standard IRC operations, Archive Team participants likewise designate their own *handles* – or nicknames, sometimes real names, sometimes pseudonyms (sometimes multiple) – by which they connect to IRC servers and communicate across a variety of other software platforms central to Archive Team activities (e.g. GitHub, the Tracker and the wiki - discussed below).

The distributed DIY nature of IRC, which requires individuals (operators) to self-host IRC server nodes, simultaneously makes it a free and open communication tool (where technically anyone with an Internet connection can connect) and one that feels simultaneously obscured by the technical knowledge and expertise required to

⁷I have deliberately chosen the term geek here, which draws on Kelty (2008, pp.35-36), where the term geek ‘signals a mode of thinking and working’. Like in Kelty’s (2008) work with free and open source software (F/OSS) communities, Archive Team includes a mix of participants (from a variety of professions and backgrounds) that are united in a central overriding concern for the preservation of information access on the Web.

engage with the network. The IRC environment reflects the look and feel of early online forums and chat rooms, where the nature of synchronous communications requires participants to either be online to receive notifications or alternatively set-up automated scripts (e.g. an *IRC bouncer*) to persistently log channel communications when users are technically offline. Since 2011, one Archive Team member has self-hosted Archive Team IRC logs of the main public channels online; an activity that was once fondly referred to by another member as A.B.L. (Archive Team *Always Be Logging*). The public chat logs are symptomatic of a widely-observed tendency for geeks and hackers to self-archive (Kelty, 2008), but practically speaking, the logs offer a mechanism for Archive Team members (old and new) to organise and search previous projects, discussions and activities over time. Snippets of IRC conversations make their way into the records of other digital transactions by Archive Team members, where logs are often quoted to provide additional context for Archive Team activities and decisions in sites such as the wiki and GitHub code README files.

The wiki also works to create organisational norms that reinforce the values of the community, by transmitting information to and enrolling new members, creating and dividing labour roles for new projects and documenting the history of Archive Team itself. Acting as the main website for Archive Team, the wiki is the primary landing page for anyone interested in their activities and although people are generally helpful in the IRC channels, new-comers with questions are actively encouraged to ‘RTFWiki’ (*Read the Fucking Wiki*)⁸ before expecting assistance from the team of volunteers. Here, the wiki as a form of self-archiving promotes meritocratic values and enables self-reliance in a way that demonstrates Archive Team’s commitment to encouraging those with the technical ability to contribute to the collective goals.

The wiki is password-protected and in a ritual reminiscent of role-playing games, Archive Team requires prospective wiki editors to first connect to IRC and ask: ‘*WHAT FORSOOTH, PRITHEE TELL ME THE SECRET WORD*’ – to which someone eventually replies: ‘*What is your quest?*’ and messages the password.⁹ Here, the gamification of receiving the password can be thought of as a routine activity that is both practically and symbolically a path to participation in the community. It signifies the transformation of newcomers (outsiders) into members of Archive Team (insiders), but also highlights the presence of a social hierarchy that distinguishes between newcomers and ‘core members’ who participate on a regular basis, have admin privileges and typically run some form of infrastructure for the group (e.g. ArchiveBot or staging servers). This social hierarchy will be revisited later in the context of the Tumblr project, where it played a role in both enabling and constraining how web archiving was enacted.

⁸This is an adaptation of RTFM (*Read the Fucking Manual*), a commonly used acronym amongst geeks and online communities wishing to express frustration with people who ask questions before first seeking out the answers themselves. See also: LMGTFY (*Let Me Google That For You*).

⁹I am told this is a Monty Python reference, as well.

The wiki also assists in the task of keeping track of dead and dying sites (discussed in Section 5.3.1) and individual project pages. Each major project gets its own wiki page which provides documentation (with varying degrees of completion) regarding the basics of the project (see Figure 5.2 for an example). Project wiki pages provide information on: the site URL, status of the project, URLs to GitHub code repositories used in the project, names of the project lead and IRC channel (if applicable), a basic history of the site in question (typically including externally-linked citations for further information), details about ways for volunteers to get involved, and notes on the processes undertaken to archive the project. Catalogues of early and ongoing project pages also act as an ‘organisational’ record for the outcomes of Archive Team Warrior projects, where templates provide information on their dedicated communication channels, start/end dates and the outcomes of projects (Figure 5.3). The success of a given Warrior project is classified by one the three ratings: *Success*, *Qualified Success* and *Failure* (Table 5.1).¹⁰ The ways that Archive Team categorises and records the success of Warrior projects is indicative of value placed in both the perception of ‘completeness’ of site archives and the archive’s public accessibility online.

Classification	Description
Success	All data captured and made available
Qualified Success	Data incomplete or the archive could not be made publicly available
Failure	The site closed before it could be archived

TABLE 5.1: Archive Team Warrior project categories of success.

Netiquette as community practice

Both the IRC channel(s) and the wiki have subsequently become the pivotal organising sites for the distributed work of Archive Team by connecting people with project documentation. Since the start of the collective in 2009, the single IRC channel has expanded to include many channels dedicated to specific platform and project communications, as well as to signal organisational norms for ‘on and off-topic’ archival-related chat. According to the wiki (and witnessed during these observations) the group and channel operators actively enforce adherence to the general rules of IRC etiquette or *netiquette* – encouraging users to ‘be helpful’, ask informed and direct

¹⁰https://www.archiveteam.org/index.php?title=Projects#Warrior_projects (visited on 6th Mar. 2019)

¹¹<https://www.archiveteam.org/index.php?title=Tumblr> (visited on 25th Feb. 2019)

¹²https://www.archiveteam.org/index.php?title=Projects#Warrior_projects (visited on 6th Mar. 2019)

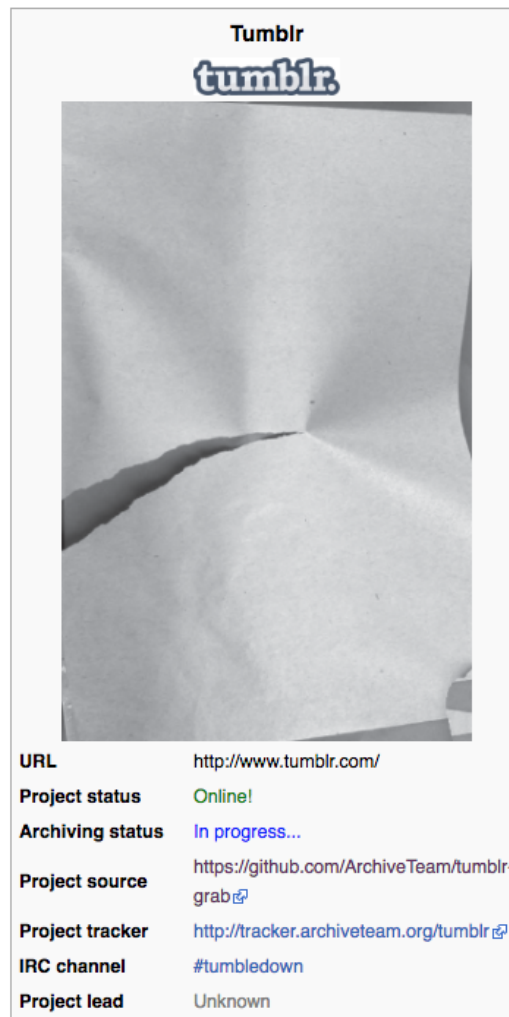


FIGURE 5.2: A screenshot from the Archive Team wiki showing basic information associated with the Tumblr project.¹¹

5.2. Framing Archive Team, Constructing Community

Warrior projects

ArchiveTeam's past, current and future Warrior projects with details, in a table form.

[Edit this list](#)

Project	IRC channel	Status	Began	Finished	Result	Archive Location
Flickr	#flickrkr	In Development				
Tumblr	#tumbledown	Active	December 8, 2018			
NUjj		Active	August 25, 2016			
Yahoo! Answers	#noanswers	Active	August 21, 2016			
Orkut	#throatkut	Active	August 6, 2016			archive
Portalgraphics.net	#archiveteam	Archive Posted	July 23, 2016	July 27, 2016	Success	archive
DNS History	#greatlookup	Aborted	July 4, 2016	August 22, 2016	Failure	
THOMAS		Archive Posted	July 3, 2016	July 5, 2016	Qualified Success	archive
Coursera	#cursera	Archive Posted	June 26, 2016	June 30, 2016	Success	archive
Olympe		Downloads Finished	June 5, 2016	June 6, 2016	Qualified Success	
ZippCast		Archive Posted	June 3, 2016	June 10, 2016	Qualified Success	archive
Arto		Archive Posted	May 8, 2016	June 29, 2016	Success	archive
Bayimg		Active	April 28, 2016			
PDF 2016	#pdfush	Active	April 8, 2016			archive
Virgin Media	#virginsacrifice	Downloads Finished	March 30, 2016	April 28, 2016	Qualified Success	
LiveJournal	#recordedjournal	Active	March 12, 2016			
GameTrailers	#unhitchedtrailer	Archive Posted	February 9, 2016	February 18, 2016	Qualified Success	archive
Fotolog.com	#fotologout	Active	February 8, 2016			archive
Friends Reunited	#friendsunited	Archive Posted	February 5, 2016	February 26, 2016	Qualified Success	archive
myVIP (script-only)	#byevip	Archive Posted	January 24, 2016	August 30, 2016	Success	archive
MusicBrainz (external links)		Archive Posted	January 8, 2016	January 9, 2016	Success	archive
OldFriends		Archive Posted	December 29, 2015	January 20, 2016	Success	archive
Google Code	#googlecodeblue	Active	December 18, 2015			archive
Docstoc	#docstop	Archive Posted	November 24, 2015	December 1, 2015	Qualified Success	archive

FIGURE 5.3: A screenshot from the Archive Team wiki showing a selection of Warrior Projects, organised by the most recent project.¹²

questions, avoid feeding the ‘trolls’ and above all else, to stay on-topic.¹³ Additional ‘Special Archive Team IRC rules’ have also emerged, reinforcing the importance of staying on-topic, but also explicitly discouraging participants from criticising or largely questioning the underlying premise of Archive Team’s archiving methods. Quoting directly from the wiki:

Don’t maliciously/demandingly criticize Archive Team, its members, nor the Internet Archive, especially in general, empty phrases. If you have a remark/idea, be concrete and constructive (and polite and patient), and if you can, realize it yourself (we’re volunteers otherwise busy). Remember the money-back guarantee!

Don’t try to convince Archive Team about that archiving is bad. We make very few exceptions when it’s about archiving. Also, our rule of thumb is ‘archive first, ask questions later’. Our IRC channels are the #1 worst place to ask ‘why we are keeping this’!¹⁴

¹³Netiquette, or ‘network etiquette’ comes in many forms, for example, see Hambridge’s (1995) RFC for *Netiquette Guidelines* that outlined the ways to keep networked (Internet) interactions polite in the absence of typical in-person communication strategies. See also Reid (1996, p.398) for a further discussion of the ‘shared systems of interpretation’ that underlie the rules of netiquette in early IRC communications.

¹⁴https://www.archiveteam.org/index.php?title=IRC#Special_ArchiveTeam_IRC_rules (visited on 4th Feb. 2019)

The introduction of the #archiveteam-bs channel for all ‘non-emergency’ archiving-related communications, as well as #archiveteam-ot for off-topic chat, subsequently made the #archiveteam channel what has been referred to as a ‘bat signal’ for rallying the Archive Team to action. Describing the difference between the #archiveteam and #archiveteam-bs channels to a new-comer, austin explains:

```
<blake> What is this channel for compared to -bs ?  
<austin> This is a channel of action.  
  
This is a channel for when the archive team is needed, for when something has  
arisen, or an issue needs to be handled.  
  
<blake> Oh... My bad.  
  
<austin> Like a firehouse, when there's no fire, people play cards and bullshit  
a lot.  
  
So then we made -bs so all the nerds can go discuss what ifs and what fors.  
  
<cameron> aka Batlight
```

Whilst further entrenching the fire metaphor, austin's description of #archiveteam-bs also intimates that in addition to being a place where work gets done, the IRC is (inevitably) a place for people with an affinity for archiving to hangout, connect and ‘bullshit’. Nonetheless, when IRC channel participants stray off-topic, channel operators (the core members responsible for informally moderating the IRC channels) will often remind participants to take conversations elsewhere. On one occasion, when a new-comer questioned why the Archive Team channels weren't officially moderated (e.g. restricted only to certain privileged users), several members responded by saying that Archive Team ‘is an anarchist organisation’ where the entry-bar to participation should be low. However, IRC infractions – for example, asking too many questions, ‘being childish’, ‘annoying’, and/or repeatedly questioning the work of Archive Team – will get channel users banned by Archive Team operators.¹⁵ For returning/repeat participants, these IRC rules work to identify and legitimise a shared notion of membership within this distributed virtual environment (Nocera, 2002), where Archive Team (re-)produces an online community dedicated to the mission of archiving web content, no questions asked.

Tenets of radical web archiving

In a podcast interview Scott likened Archive Team's mode of web archiving to that of Earth First! (Findlay, 2011), a radical environmental activist coalition known for their preference for direct action and civil disobedience (Lee, 1995). I want to briefly explore this comparison, as it illuminates several aspects of the ways that Archive

¹⁵The IRC wiki page outlines both the general and specific rules of channel communications for Archive Team, as well as provides links to examples of prohibited behaviour, illustrated in IRC logs hosted elsewhere. See the wiki notes for examples of actual infractions: https://www.archiveteam.org/index.php?title=IRC#Special_ArchiveTeam_IRC_rules (visited on 4th Feb. 2019)

Team positions itself (both explicitly and implicitly) in opposition to the forces of ephemerality and those that might question or resist their methods.

In the podcast, Scott contrasts the tactics of Earth First! with those of the Sierra Club (another environmental advocacy group) distinguishing between ‘covert’ and ‘legitimate’ (or what may be considered ‘mainstream’) forms of environmental activism, and signals Archive Team’s preference for direct action and oppositional forms of resistance (Findlay, 2011). Whilst possibly being in danger of extending this analogy beyond the irony clearly intended, exploring the attributes that make Earth First! ‘radical’ is instructive for situating Archive Team in relation to their mainstream counterparts. Lee (1995) describes the core tenets of Earth First! that are particularly salient to a discussion of Archive Team activities as a form of ‘radical’ web archiving:

“Most often, they confront environmental problems through direct action (and might be willing to destroy property); the goal of their protests is the preservation of biological diversity; they act without direction from an organizational hierarchy; they are poor; and they have little hope of actually ending the practices against which they protest” (Lee, 1995, p.9).

This highlights several analogous tenets of practice that are observable in the ways that Archive Team frames and enacts web archiving. Here, the work of Archive Team is centred around the value of direct action to intervene in the closure of websites and web services through the act of web archiving. As a result of their adherence to a ‘brute force’ approach to web archiving they routinely (and unintentionally) create the equivalent of ‘denial-of-service’ (DoS) attacks when the number of participants outweighs the bandwidth or capabilities of the targeted site. Archive Team espouses a commitment to ‘biological diversity’ by collecting any and all websites, and as drew reflected, they work without any illusions of stopping the Web’s ephemerality or indeed website closures. The following considers and expands on the above through two tenets of practice that are frequently used to frame the work of Archive Team.

Brute Force Archiving At every opportunity Archive Team makes it clear that they largely do not ask permission to archive the Web. The question of seeking permission is frequently raised by new participants and interested parties in the IRC channels and in public forum discussions about the work of Archive Team elsewhere online. More than a trivial question of process, the issue of permissions is indicative of a fundamental difference between Archive Team, conventional (institutional) web archiving efforts and an emerging class of Internet researchers who continue to question web archiving collection activities in the absence of informed consent from content creators.¹⁶ For better or worse, this action-oriented ‘get things done’ approach has enabled Archive Team to proceed where institutional web archives (such as national web archives) have been subject to their own institutional mandates and policies (e.g.

¹⁶For example, see the Documenting the Now Project: <http://docnow.io>

legal deposit schemes) that restrict the nature of what can be collected, stored and made accessible to the public.

However, drew explained that sometimes they do approach site owners, often in an effort to facilitate smoother or more comprehensive archiving of the site in question:

“Sometimes people – and I mean, I understand and it’s their business and they decided to shut the website down – and if you see the list of websites that are being archived by ArchiveBot it’s pretty hard to ask each owner of the website for permission. And then I’m pretty sure in 75% of the cases they won’t give permission and then we’re still archiving it anyway - so we’re not really asking permission. Sometimes we’re sending an email upfront with a question - if we can have a list of users from the website for example. Yea I guess we’re not really polite. [...] So yea they can get a little angry.” (drew)

Scott has argued that Archive Team’s approach ultimately operates to ‘keep the discussion going’ even if (at times) their tactics are internally recognised as imperfect and haphazard (Findlay, 2011). Archive Team frequently takes an ‘archive first, ask questions later’ strategy which emphasises action over discussion of best practices – more than implying that other practitioners in the libraries/archives field are frequently blocked by philosophical debates about practice while ‘the building is on fire’.¹⁷ This is best summarised in the Archive Team’s decisions to disregard the `robots.txt` protocol, which was made particularly clear through Scott’s deliberately provocative post on the Archive Team wiki entitled: ‘Robots.txt is a suicide note’ (Scott, 2017). Despite the aims of keeping the discussion going, Archive Team’s brute force approach has worked to both open and close the lines of communication between themselves, platform-services and users. For example, following Archive Team’s efforts to archive DNSHistory – an online archive of historical DNS records – put up a permanent banner notification that reads:

“DNS History has now shutdown, any updates are for my personal requirements only – the site may go up/down as I need the resources it uses. The number of servers dedicated to the DNS History project is now ZERO. Due to the Archive Teams self-righteous attitude CloudFlare’s DDoS protection has been enabled. They are quite open about their attitude of ignoring robots.txt files and work around blocks on their user agent – this [is] abuse.”¹⁸

Despite the objections of sites like DNSHistory, in the spirit of radical web archiving, Archive Team and the image they project may be akin to a contemporary form of what Hobsbawm (1959; 1972) describes as ‘social banditry’. Social banditry is characterised as a ‘primitive form of organized social protest’ (Blok, 1972, p.494) enacted

¹⁷‘Archive first, ask questions later’ is listed in the Archive Team IRC rules described above in Section 5.2.2.

¹⁸<https://dnshistory.org> (visited on 9th Mar. 2019)

by outsiders who reject societal norms to plunder and redistribute their wealth to the poor and vulnerable.¹⁹ In this analogy, Archive Team is positioned so that the work of web archiving becomes a necessary force for good, despite objections from others that may be opposed to their methods. Whilst I want to avoid getting sidetracked with a thorough discussion of the decades of critique to this particular social theory, the analogy nonetheless provides another window into how Archive Team positions its own form of citizen action as a moral good to in effect, ‘save it for the people’ (Scott, 2009a).

Archival Neutrality Just as the Earth First! collective has a basic tenet of equal respect for all living creatures (considered radical by some) – so too does Archive Team espouse to treat all websites with equal priority.²⁰ Although questions surrounding the power and politics of archival selection have historically preoccupied the profession for decades, it is clear that Archive Team generally and wholeheartedly resist any notion of succumbing to these particular ‘flights of fancy’.²¹ When I asked about how sites get selected and prioritised, drew shared with me this passage from Archive Team’s Wikipedia page, followed by their own interpretation:

“According to Jason Scott, ‘Archive Team was started out of anger and a feeling of powerlessness, this feeling that we were letting companies decide for us what was going to survive and what was going to die’. Scott continues, ‘it’s not our job to figure out what’s valuable, to figure out what’s meaningful. We work by three virtues: rage, paranoia and kleptomania’.”²²

“I think we should just archive as much as possible. From any kind of thing. If it’s a website with people posting a daily image of their sandwich and that website is going offline I think we should archive it. Yea we can’t know now what will be important in a hundred years. I mean these more extremist people will only want to archive websites that are more like how they think about the world and that’s one way of archiving but it’s not a way of archiving that makes sure we have a good view later on of how the world was. So I think we should not make any decisions – or make as little as possible decisions on what to archive. As long as it’s going offline and it’s accessible and it’s not illegal to have in your possession then I think we should be able to archive it.” (drew)

¹⁹Elsewhere Coleman (2014, p.71) has also compared and contrasted the tactics of *Anonymous* to Hobsbawm’s concept of social bandits.

²⁰Earth First! subscribes to the notion of *biocentric equality*, or recognition of ‘the intrinsic moral worth of both human and non-human life’, which drives a view that all life is sacred, interconnected and worth saving (Lee, 1995, p.18).

²¹See Chapter 6 for further discussion about the politics of archiving.

²²This is quoted from the Wikipedia page here: https://en.wikipedia.org/wiki/Archive_Team (visited on 5th Jul. 2019). The quotes were attributed to snippets of Jason Scott’s keynote at the 2012 Open Source Bridge conference (Scott, 2012).

The Wikipedia passage (drawn from a presentation by Scott [2012]) again positions Archive Team's projects in opposition to corporations and platforms that shutter their services without warning to users, emphasising their role in deciding 'what [is] going to survive'. By implication, Scott positions the role of Archive Team through a lens of objectivity, championing a general neutrality towards deciding the value of collecting certain websites over others. This view is extended by drew's reflection that collection must be representative of the diversity of online experience so that in future 'we have a good view' of 'how the world was'. Each of these reflections are rooted in the value of neutrality in archival creation, and work to provide context to the ways Archive Team deploys the project of web archiving in pursuit of saving the Web's history.

These two tenets of practice – *brute force archiving* and *archival neutrality* – work to frame an approach to web archiving that is informed by the cultural components of the community espoused in previous sections. From a culture as practice point of view, these tenets form the basis to which Archive Team members and participants can look to deploy strategies for web archiving in practice. However, tensions in each were observed through the Tumblr project, where on the one hand, (inevitable) selection decisions were both observed and challenged through the ad-hoc decision and consensus-making required to meet the challenges of the project. Similarly, the brute force approach produced a set of circumstances that revealed the strategies of action deployed in the face of repeated attempts by Tumblr to ban Archive Team from crawling the site. Here, as will be discussed, new participants relied on the conventions of practice (as dictated by core members) both to enact but also challenge ways of working that ultimately transformed the nature of what was archived in practice.

5.3 Archive Team at Work

5.3.1 Yahoo and the Case of Tumblr

The story of Archive Team's engagement with Tumblr begins in 2009 with their early archival interventions with other Yahoo-sanctioned closures. The shuttering of platforms such as GeoCities (acquired by Yahoo in 1999) – and the archiving project that followed – marked both Archive Team's earliest major archival intervention, and the first of many dealings with Yahoo platforms and services. Since early 2009, Yahoo services and platforms have been at the top of Archive Team's *Deathwatch*, with the collective 'officially [proclaiming] 'Yahoo! the least trustable host and its arch-enemy'.²³ And perhaps true to form, in the case of Tumblr, this culminated in a flurry of activity just before, during and after the Christmas holiday period of 2018, with Tumblr's announcement giving Archive Team just two weeks to archive NSFW blogs.

²³<https://www.archiveteam.org/index.php?title=Deathwatch> (visited on 12th Jan. 2019)

According to the Archive Team wiki, Yahoo has had a history of closing platforms with little-to-no notice or provision for mechanisms for users to retrieve their content – creating a deep sense of distrust in the longevity of any platforms or domains originating from or acquired by Yahoo and its subsidiaries.^{24,25} Speaking about some of Yahoo’s ‘victims’, Scott outlines their frustrations:

“These are all the things that Yahoo! has shut down in the last four years. Just so you understand. Yahoo briefcase, where you were able to store 10 megs, whenever you wanted, and get it from anywhere via FTP. They shut down. Why? No spare USB drive? Content Mash, some of these you won’t know.

Yahoo Pets was funded by Purina for a five year contract and on the day that the contract ran out they shut it down and redirected it to Yahoo Women. I don’t know why, but they did.

But it was a case of there was this secret contract, and when I say they shut down, I mean with no warning. One day it was there, one day it was gone. It had pet pictures, it had forums in it, everything, gone. Totally gone. So in other words I’m saying Yahoo blows, OK? It is a fucking clown car. I wouldn’t trust them with like a backup of my nutsack, because these guys... This is a case where a company went speculatively into user generated content and when they decided it wasn’t worth it any more, they got out of it” (Scott, 2011).

This sheds light on several aspects of the project of archiving Tumblr, but also some motivations that underpin the wider work of Archive Team. As portrayed above, Archive Team’s grievances with Yahoo stem from a pattern of disregard for users’ rights to control what happens to the platform and their contributed content – particularly during precarious points of platform governance (such as when external services are procured, taken-over, purchased and transferred). Here, Yahoo is accused of short and unreasonable timelines for platform closures which prevent users from retrieving their content contributions to the sites in question. Problems of platform sustainability are not limited to Yahoo of course, and the scale and breadth of projects undertaken by Archive Team demonstrates both the patterns of online closures, but also the multiplicity inherent in the root causes of platform retirement and consequentially, the ephemerality of user-generated content online.²⁶

And so, by the time Tumblr announced on the 3rd of December, 2018 that it would be taking steps to permanently block ‘adult content’ on the platform from public

²⁴<https://www.archiveteam.org/index.php?title=Yahoo!> (visited on 12th Jan. 2019)

²⁵<https://www.archiveteam.org/index.php?title=Wooohoo> (visited on 12th Jan. 2019)

²⁶Archive Team has even outlined their ‘warning signs’ of pending closures, attributing site closures to a multitude of underlying causes and guiding users to be wary, vigilant and of course, take back-ups. See: https://www.archiveteam.org/index.php?title=Warning_Signs (visited on 2nd Mar. 2019)

view, Archive Team took immediate notice. Since its launch in 2007, Tumblr has become a widely used micro-blogging platform with an estimated 455 million blogs and over 168 billion posts at the time of writing.²⁷ Tumblr's content moderation policies have hitherto been considered relatively permissive, particularly in comparison to other popular social media platforms like Facebook and Instagram. Tumblr had become a space for the curation and circulation of 'adult content', including manga and user-generated fan art (art based on popular works of fiction) which is erotic and sexually-themed in nature. Policies, in combination with other sociotechnical platform affordances have worked to support 'a sense of community' and belonging amongst users/tumblelogs (blogs) dedicated to 'counterpublics' considered fringe and marginalised IRL ('in real life') or elsewhere online (Cho, 2015a; Hart, 2015; Tiidenberg, 2014, 2016). As researchers have shown, this is substantiated by the emergence and growth of tumblelogs dedicated to supporting networks of queer, transgender, transsexual and gender nonconforming communities (Cho, 2015a,b; Fink and Miller, 2014), a-sexual (Renninger, 2015) and polyamorous communities (Tiidenberg, 2014), curation cultures surrounding self-injury and recovery (Seko and Lewis, 2018), as well as those negotiating youth intimacy and social isolation (Hart, 2015).

After Yahoo acquired Tumblr in May 2013, concerns were raised by platform users (e.g. see reactions in Renninger, 2015, pp.1519-1520), researchers (e.g. see Fink and Miller, 2014), tech journalists (Lunden, 2013) and Archive Team alike. Following Tumblr's purchase, they were all united in their scepticism that then Yahoo CEO Marissa Mayer – regularly cited as an adversary by Archive Team – was actually capable of keeping their promise to 'not screw it up' (Mayer, 2013), particularly in the face of pressure to generate increased advertisement revenue from the platform. Between 2013-2018, Yahoo and subsequently, Verizon,²⁸ took a series of steps to moderate content on the platform through both user-based and algorithmic filtering; an effort widely interpreted to be in response to both a desire to boost advertisement sales and increased pressures from the Apple app store to censor pornographic content (Gillespie, 2018, p.175). Tumblr allowed users to self-tag their blogs as *Not Safe For Work* (NSFW), creating a hybrid browsing experience distinctly different from pornography, routinely '[alternating] indiscriminately between posts of cupcakes, fashion, kittens, and cocks of the flesh, synthetic, and illustrated varieties' (Fink and Miller, 2014). Tumblr's choice to not distinguish between the boundaries of (for example) 'artistic, casual or pornographic nudity' (Gillespie, 2013) created a problematic environment for both increasing ad sales, and meeting the terms and conditions of Apple's iOS app guidelines for filtering content.

When Tumblr introduced the 'safe mode' in June 2017 and turned it on site-wide

²⁷<http://web.archive.org/web/20190112191359/https://www.tumblr.com/about> (visited on 14th Jan. 2019)

²⁸Verizon Communications purchased Yahoo's Internet business for \$4.48 billion (USD) in June 2017 (Goel, 2017).

in February 2018 (effectively blocking all NSFW posts from public view by default without a login),²⁹ the Tumblr universe reacted, including claims that ‘art and culture were getting destroyed’ (Koebler and Cole, 2018). However, the identification of child pornography on Tumblr in November 2018 was the final straw. Apple subsequently removed the application from their store, and Tumblr started rolling out a series of interventions to ‘algorithmically’ identify and block content it now deemed pornographic (Koebler and Cole, 2018). NSFW posts were among the first to go.

Prior to 2018, Archive Team had, in fact, already made several previous attempts to archive Tumblr – first using the platform as a test case for their new version of the Warrior application in 2012, and again after Yahoo placed Tumblr in their ‘decide on’ category in a 2015 shareholder presentation (Spring Owl Asset Management LLC, 2015). After Tumblr was purchased by Yahoo in 2013, Archive Team members were concerned about the platform’s eminent demise, but many expressed the opinion that the job was ‘too big’ and that there wasn’t enough space to cope with the size of Tumblr and its posts – estimating it was at least 28 petabytes. The 2015 #archive-team IRC logs confirm that several members had actually been using ArchiveBot to archive Tumblr posts ‘for a couple of years’. Although the Yahoo shareholders presentation led to the creation of the #stumblr IRC channel in 2016, as there was little to no documentation on the wiki about the channel or this round of activities, with 2017 came a renewed interest in Tumblr and along with it, a new project and IRC channel: #tumbledown.

This sequence of disparate undertakings is indicative of one of the major challenges of organising the distributed work of Archive Team. With dynamic and often fleeting participation in Archive Team activities by various stakeholders, the wiki became an even more central component in documenting the ‘institutional memory’ of Archive Team itself, as well as the challenges and results encountered during its previous exploits. Archive Team’s attempts to archive Tumblr’s NSFW blogs in 2018 would face many of the same difficulties encountered in their previous attempts to archive this and other platforms.

5.3.2 Distributed DIY Web Archiving

On the 8th of December, Archive Team launched the public Tracker for Tumblr NSFW blogs,³⁰ allowing those using the Warrior application to begin contributing to the project. After a week of testing the scripts, the Tracker was publicised and members used various media outlets to attract participants for archiving Tumblr NSFW. Since 2009, Archive Team has been archiving and depositing web archives at the Internet Archive, as a matter of course. The long-term storage of Archive Team projects is

²⁹<http://web.archive.org/web/20170629043336/https://staff.tumblr.com/post/162047130530/safe-mode> (archived 29th Jun. 2017; visited on 28th Feb. 2019)

³⁰<https://tracker.archive-team.org/tumblr> (visited on 2 Apr. 2019)

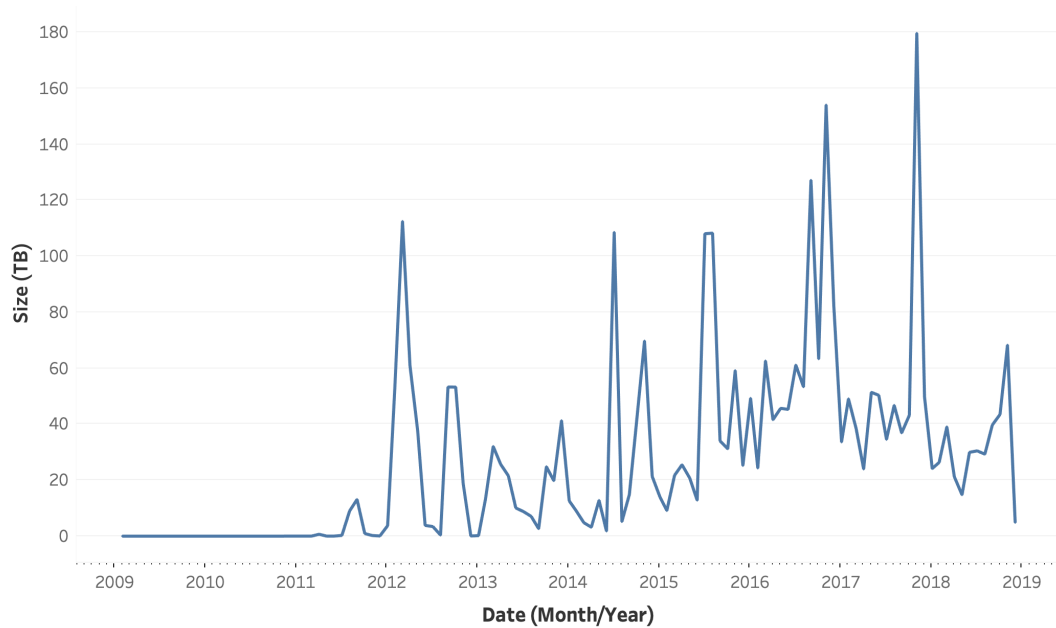


FIGURE 5.4: The rate of Archive Team collection activities between 2009 - 2018, in terabytes over time.³²

facilitated by a close collaboration with the Internet Archive; one which is manifested in both an overlap of participants³¹ and a mutually beneficial relationship whereby the coverage of the Wayback Machine is significantly extended by acting as a reliable, long-term repository for Archive Team. The scale of Archive Team's contribution to the Wayback Machine can be seen from the Internet Archive's Scrape API, where as of January 2019, Archive Team has contributed more than 9.3 petabytes to the Archive Team Collection (Figure 5.4).

Archive Team archival efforts are typically driven by early decisions from core members about which scripts and tools will be used for which projects, determining which sites require and constitute a large-scale project and which jobs can be handled on an ad-hoc basis. In a 2011 podcast Jason Scott referred to this process as 'researching the victim' to understand the intricacies of different targeted infrastructures (Findlay, 2011), again using humour and irony to provocatively situate corporate services on the losing end of Archive Team's efforts to avenge users' misplaced trust in free hosting and platform services.

For Tumblr, initial 'research' was done by members to ascertain the scale of the site and whether or not any access controls (e.g. logins or cookies) are required to archive

³¹This research supports the observation that various Internet Archive staff members have regularly participated in Archive Team projects and Archive Team members have gone on to be employed by the Internet Archive. Jason Scott went on to be employed by the Internet Archive in 2011 as a 'free range archivist'.

³²The underlying data for this calculation was generated using a tool developed by Ed Summers which scrapes the metadata associated with collections publicly hosted on the Internet Archive using the Archive's Scrape API. The code is hosted in a jupyter notebook here: <https://github.com/edsu/notebooks/blob/master/ArchiveTeam.ipynb>

it. If deemed an appropriate size, sites are submitted to *ArchiveBot*, a semi-automated bot for distributed web archiving.³³ *ArchiveBot* enables core Archive Team members with permissions to access and issue commands to the #archivebot IRC channel to request sites to be archived. Requests are then routed from a central ‘control node’ server to self-hosted pipelines operated by Archive Team volunteers, which then download the requested sites using *wpull*, sync to a staging server and upload them to the Archive Team collection at the Internet Archive.

Decisions about which tool to use are based on a combination of the time available before site closure and the scale of the site in question. However, in the specific context of Tumblr, several additional factors came into play. First the scale of the targeted site was deemed to be too large (an estimated 700,000 NSFW blogs) for *ArchiveBot* in the amount of time required (two weeks). Second, on the Tumblr platform, logged-out European-based users have to consent to Tumblr data terms and conditions (based on EU GDPR data protection laws) as well as the platform content-viewing restrictions of safe mode. These factors meant that in order to automate crawling, cookies had to be used to bypass the banners and settings – a function that, by default, is not available via *ArchiveBot*. These cookies added an extra layer of complexity to the crawling activities and reflects both the impact of geopolitical regulatory environments on web archiving at scale, but also the (intentional) constraints of particular tools in use by Archive Team. The cookies proved additionally cumbersome as the scale of operations (and participation) required a sufficiently wide-ranging and dynamic source of cookies (a ‘cookie jar’ or what Archive Team called a ‘cookie factory’) in the face of Tumblr’s efforts to mass-ban crawlers based on large-scale cookie exploitation.

Deploying the Warriors

The *Warrior* is a ‘virtual archiving appliance’ that when installed uses the host to crawl projects centrally managed by the collective via the Tracker (discussed below) and GitHub repository (Archive Team, n.d.[a]). The virtual machine (VM) reduced the barrier to entry for Archive Team participation, enabling users to download and install the appliance with little to no technical skills for interfacing with code. The *Warrior* application subsequently changed the way that Archive Team archives at scale. According to Scott, the photo sharing site *Tabblo*³⁴ was the ‘first real run’ of the Archive Team *Warrior*. Whereas previously it took 12 to 24 people, 4 to 5 months to download *GeoCities* using GNU *wget*,³⁵ it took 59 participants approximately 36 hours to archive *Tabblo* using the *Warrior* (Scott, 2012). Setting aside the issues with making a direct comparison of these two projects (*GeoCities* was undoubtedly a much larger site to archive), these figures (and Scott’s enthusiasm) signify a distinct turning point in the Archive Team’s ability to scale their archival operations. Beyond

³³<https://www.archiveteam.org/index.php?title=ArchiveBot> (visited on 31st Jul. 2019)

³⁴<https://www.archiveteam.org/index.php?title=Tabblo> (visited on 22nd Feb. 2019)

³⁵https://www.archiveteam.org/index.php?title=GeoCities_FAQ (visited on 22nd Feb. 2019)

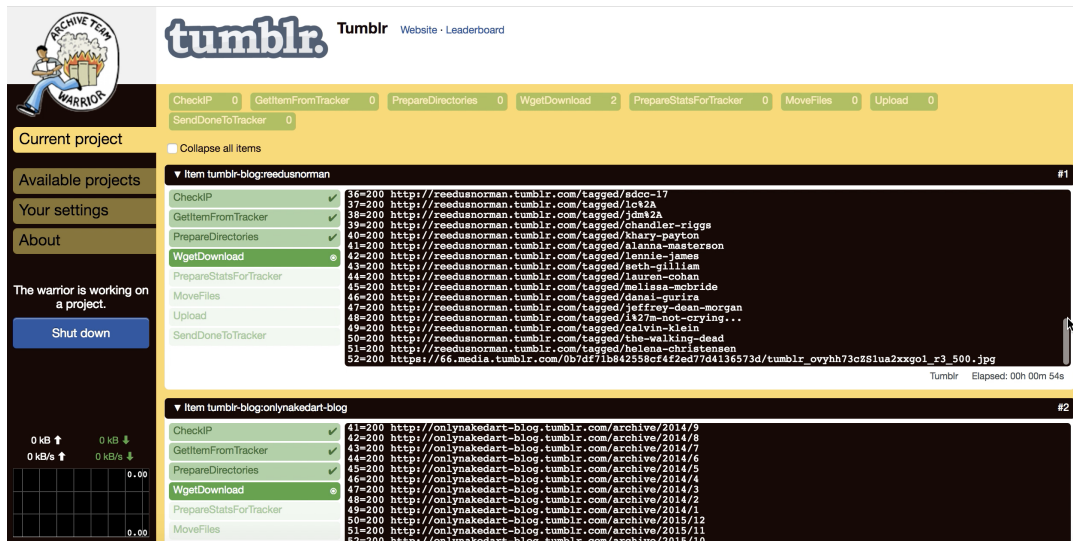


FIGURE 5.5: The Warrior application downloading Tumblr NSFW blogs.

an argument for efficiency, the Warrior – as a stand-alone software application – also signifies a concerted effort to enrol volunteers that may not have the technical expertise to administer the server environment, dependencies and scripts required to participate in distributed crawling.

Overall, the Warrior is designed for minimal interventions and no programming skill required. In order to participate in archiving, I downloaded and installed the virtual machine onto an old laptop. Within a few minutes I had the Warrior application up and running. Each Warrior is partially customisable through an interface accessible via a web browser, allowing users to select the level of *concurrency* (number of crawling threads running simultaneously) and the option to either proceed with ‘Archive Team’s choice’ or specific projects of interest. Through the Warrior I was able to monitor progress in realtime through each step of the crawl, beginning with when my machine was allocated seeds to download from the Tracker until they were uploaded to the staging server (Figure 5.5).³⁶

In addition to the Warrior (and the underlying scripts that power the distributed crawling) the *Tracker* forms a central component to the Archive Team’s web archiving infrastructure (Figure 5.6), as it controls the allocation, distribution and rate of *items* to each worker instance,³⁷ allowing web archiving efforts to work at scale. Added in late 2011, and administered by different core Archive Team members, the Tracker is responsible for handing out items to be downloaded and keeps track of whether

³⁶In fact, none of my items actually made it through the whole process as I was crawling over the wifi, which it turns out is not ideal. In the end I had to give up my attempts to use the Warrior.

³⁷The wiki explains that ‘items can be usernames, subdomains, full urls, basically any unit we can use to break the site into manageable chunks’. <https://www.archive-team.org/index.php?title=Tracker> (visited on 22nd Feb. 2019)

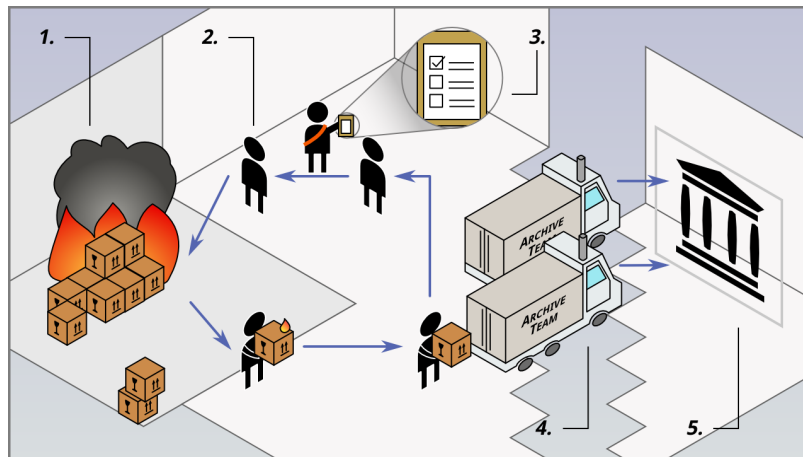


FIGURE 5.6: Graphic depiction of the Archive Team’s Warrior infrastructure, taken from their wiki. Key: 1) Website in Danger; 2) Warrior; 3) Tracker; 4) Staging Server; 5) Internet Archive (Archive Team, n.d.[b]). Image: (Foo, 2013)

or not they have been completed (e.g. uploaded to centralised staging servers).³⁸ The Tracker leaderboard publicly displays a dynamic list of user-handles running the Archive Team scripts or Warrior VM for each project, alongside the total amount of *data* uploaded (measured in gigabytes). The Tracker signifies the value of gamification in mobilising participation in web archiving, as Scott notes, ‘it turns out, leaderboards make people do things’ (Scott, 2012). The Tracker is also a window into the wider community of participation in Archive Team activities by those whose participation is otherwise not visible through IRC chat or the work of maintaining the wiki. Here, scale is measured by the number of worker handles, gigabytes downloaded and subsequently uploaded. For Tumblr, the Tracker became *the* record of participation in Archive Team archival efforts, revealing a total of 1525 handles that eventually archived c. 350,000 NSFW items (Figure 5.7).

There were persistent and sustained efforts on the part of Archive Team participants to find ways to make the Warrior infrastructure more efficient. These strategies often manifested in the collective knowledge sharing required to setup and deploy infrastructure using the Warrior scripts, as well as test the technical capabilities and limits of individual server environments and service providers for scaling Warrior deployments. Archive Team participants shared free ‘compute credits’ and information about where to find the best deals on cheap cloud infrastructure providers. They advised each other on developing efficiencies for concurrent crawling, for example the number of jobs and machines (with what RAM, disk space and projected bandwidth) maximised computing capabilities, and more importantly, archived the most in the time allotted.

³⁸The first initial commit of the universal-tracker code in the Archive Team repository is dated November 2011: <https://github.com/ArchiveTeam/universal-tracker/commit/6ae619ee7bf56bfb2cc5033c994ea9f77e6df6a3> (visited on 22nd Feb. 2019)

³⁹<http://tracker.archiveteam.org/tumblr> (visited on 3rd Jan. 2019)

You can help. [Run your own ArchiveTeam Warrior.](#)

Tumblr tracker

items	354672 done + 568938 out + 0 to do		tumblr-blog:flurry-o	13170 MB
			tumblr-blog:specspe	14563 MB
data	68878 GB	199 MB/u	tumblr-blog:pseudol	5964 MB
			tumblr-blog:hanniba	591 MB
•	5939 GB	39657 items	tumblr-blog:xchloe	13129 MB
•	4216 GB	17267 items	tumblr-blog:woah-lc	13446 MB
•	3828 GB	25816 items	tumblr-blog:twilight	11525 MB
•	3214 GB	3421 items	tumblr-blog:imthehi	5645 MB
•	3132 GB	14108 items	tumblr-blog:designe	800 MB
•	2407 GB	14965 items	tumblr-blog:suzanzi	14 MB
•	1956 GB	6088 items	tumblr-blog:adserto	6965 MB
•	1284 GB	5290 items	tumblr-blog:sheabut	7528 MB
•	1239 GB	8456 items	tumblr-blog:tainted-	2082 MB
•	955 GB	3437 items	tumblr-blog:lesbiank	13225 MB
•	873 GB	1334 items	tumblr-blog:scorpio	9179 MB
•	868 GB	1965 items	tumblr-blog:bluesky	9519 MB
•	804 GB	2229 items	tumblr-blog:last-lev	11676 MB
•	787 GB	3114 items	tumblr-blog:jennphc	9306 MB

+ Show all

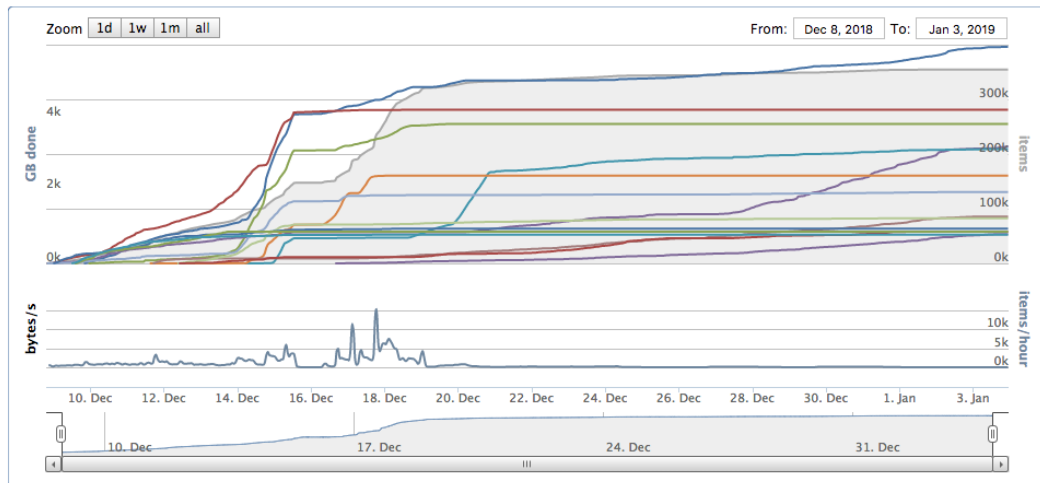


FIGURE 5.7: Screenshot from the Archive Team Tumblr NSFW Tracker, displaying the uploaders and uploads still in process.³⁹

Mobilising the Masses

Online media played an important role in mobilising and enrolling participants in Archive Team activities, including their own efforts to self-promote project-specific causes via Twitter and Reddit, as well as numerous articles written and published by journalists and media outlets (Duguay, 2018; Ho, 2018; Koebler and Cole, 2018; Kozłowska, 2018; Locker, 2018). In a strategic effort to mobilise Warriors and raise awareness of the cause, Jason Scott tweeted at various key stages of the operation (Figure 5.8), opting to solicit new participants only when the Tracker was moving speedily along. Other core members used Reddit (/r/datahoarders and /r/archiveteam) and Tumblr itself to mobilise technical expertise on the one hand, and Tumblr stakeholders in NSFW archives on the other.^{40,41} Further to bringing in new members (as a form of enrolment), the media attention also therefore becomes vital to the success of the project by scaling operations and bringing in additional resources in the form of infrastructure, labour and sociotechnical expertise.

During the course of this research, I regularly witnessed participants wandering into the IRC channels after having either recently read about Archive Team on Reddit, in a popular media article, or because their favourite platform was being sunset and they wanted to archive it. Apart from the occasional reporter, the IRC channels are usually limited to involvement from prospective Archive Team participants with a stake in the outcomes of Archive Team's efforts – to announce that a site is going down and request that it be archived, to ask how to get involved, and to more generally proclaim their approval of Archive Team activities (although in some cases, to express their disapproval). In addition to encountering Archive Team in the press, drew described the role of their own personal interests in motivating involvement, as well as the effects of a 'local' platform closure (e.g. a platform relevant to their own local use) on their decision to participate:

"I think I became involved in the way that many people are involved. I became involved I believe six or seven years ago. [...] a larger Dutch social media site was shutting down and I read this news article that said that volunteers were needed by Archive Team to save this social media website. So I was like 'hey that's pretty cool' and I didn't have much to do in my free time. So I checked it out and helped archiving the social media website and then I kind of stuck around and eventually started doing more and more in Archive Team and now I'm writing scripts for larger projects." (drew)

"I became involved in [Archive Team] sometime in June 2018 [...] [Pure Volume] was disappearing, and we had a project called puresilence. Since I had used [Pure Volume] for some of my life, I wanted to make sure that at least artist profiles that I listened to was preserved in some way." (ezra)

⁴⁰<https://www.reddit.com/r/datahoarders> (visited on 31st Jul. 2019)

⁴¹<https://www.reddit.com/r/archiveteam> (visited on 31st Jul. 2019)



FIGURE 5.8: Screenshot of Jason Scott tweets used to mobilise participants in Archive Team's efforts to archive Tumblr NSFW posts.

Both drew and ezra point towards the closure of two social media sites as the motivations for getting involved. Despite the numerous challenges encountered whilst archiving Tumblr, volunteers continuously showed up during the two weeks in #tumble-down and enquired about how to get involved. This can at least be partially attributed to the use and publication of media updates during the course of archiving Tumblr NSFW, where an influx of participants would enter the IRC following big announcements either on social media or by mainstream reporting on the efforts.

Other efforts by Archive Team to enrol stakeholders included the creation of an open Google Form to enable public seed submissions. The form was promoted on social media (including on Tumblr), demonstrating the ways that some members attempted to channel support for the cause by capitalising on the mounting Tumblr backlash using associated trending hashtags (e.g. #dec17, #december17) (Figure 5.9). The Archive Team Tumblr posts illustrate efforts by members to disseminate information about how Archive Team works in practice, and solicit support through re-blogs, including invitations to participate in the project on IRC and to donate to the Internet Archive.

5.3.3 Transforming Culture

Despite their idealistic stance towards selection discussed earlier – or what I’m calling an ideal commitment to *archival neutrality* – observations support the notion that Archive Team members are, in fact (unavoidably) making selection decisions during the course of their archiving activities. In general terms, through the process of ‘researching the victim’, selection operates on at least three levels where choices are made about what projects are undertaken, what seeds and URLs are archived and how particular components of sites and platforms are included and excluded from the archives. This section outlines the ways in which the decisions concerning *what gets archived* are informed by and contingent on the time and resources available, the value-judgements and priorities of a dynamic combination of stakeholders (Archive Team participants, the Internet Archive, and ‘the crowd’) and the sociotechnical affordances of the targeted platform and tools chosen to undertake the work. By observing practice in action, strategies of action are revealed through dilemmas that evidence the impact of a discourse of abundance, various ‘folk theories’ about platform functionality and value claims surrounding decisions to scope the collection.

Archival Neutrality and Selection in Practice

On one level, collective selection decisions are made surrounding what projects the Archive Team undertakes and in what ways the project is carried out. It is undeniable that in the round, Archive Team is relatively open to archiving just about anything on the public Web. As briefly discussed, Archive Team regularly employs a range of

Save endangered tumblrs before 17. December

Please submit all endangered Tumblrs you care about [in this form](#). Also please spread the message.

Hello Tumblr users,

we are the Archive Team, the Archive Team is a loose collective of rogue archivists, programmers, writers and loudmouths dedicated to saving our digital heritage. Since 2009 this variant force of nature has caught wind of shutdowns, shutoffs, mergers, and plain old deletions - and done our best to save the history before it's lost forever. We already worked on other Yahoo related projects like GeoCities and Flickr.

You can learn more about us under <https://archiveteam.org/>.

We plan to save many as many of the endangered blogs as possible. To do that we take multiple approaches.

1. We use or scan of Tumblr we did in April 2018
2. Some guys are working on an AI solution to detect endangered pictures
3. We look for tags manually

But just us looking is not enough. You know Tumblr better than us.

Submit your favorite endangered Tumblrs [here](#). That way we can be sure not miss them and put them in front of our queue.

If you have technical knowledge join us via IRC chat.efnet.org:9090/?channels=%23tumbledown

#tumbledown #17 december #december 17 #archives #censura
#new rules

2,266 notes Dec 4th, 2018

(A)

<https://17decemberarchive.tumblr.com/post/180795765906/save-endangered-tumblrs-before-17-december> (visited on 9th Mar. 2019)

Everyone on tumblr on the morning of December 17th:



 17decemberarchive

We, the archive team, think that **petitions will change nothing**. That is why we are preparing for the worst. We have dealt with Yahoo before. They killed GeoCities with little warning ([read more](#)). Currently Flickr another Yahoo subsidiary is on the death bed too. That is because all accounts which have more than 1000 pictures will be frozen and many pictures deleted until 1000 remain. We do not have high hopes that Yahoo will turn around on this one issue of many. Do not get us started on [their Blogs, Video Platform and Groups](#).

We attempt to archive most of the NSFW blogs of tumblr. Instead of singing partition you can [back up your own data](#) and [tell us about your favorite endangered blogs so we can archive them](#).

The tag of the campaign is #tumbledown btw. You can learn more about us here:

<https://17decemberarchive.tumblr.com/post/180795765906/save-endangered-tumblrs-before-17-december>

#17 december #petition #archives #archivist #archive team
#tumbledown

8,700 notes Dec 5th, 2018

(B)

<https://17decemberarchive.tumblr.com/post/180816781477/we-the-archive-team-think-that-petitions-will> (visited on 3rd Mar. 2019)

FIGURE 5.9: Two posts that Archive Team promoted on Tumblr to generate submissions and support for the web archiving project.

mechanisms for monitoring and soliciting nominations for sites that are in need of archiving – through the use of IRC and social media. As will be discussed further, selection is based on a myriad of personal and crowd-based mechanisms for realising sites are dying. Often spurred by media announcements pertaining to platform closures, people request and suggest sites to archive on IRC, Twitter, Reddit and through email and direct messages to known Archive Team members. Some Archive Team members setup Google alerts, and monitor subreddits like r/shutdown, a subreddit dedicated to announcing the demise of start-ups and other web media.⁴² Additional mechanisms for nominating platforms and specific content therein have been implemented in other projects, including the use of bots which listen to specific hashtags on Twitter⁴³ and Google Forms for individual nominations (as was used in Tumblr, discussed later). Other IRC bots are used to ‘listen’ to additions to Wikipedia categories such as *Deaths* and *Disestablishments*, which provide sources for an automated pipeline that adds the official website for each entry to a list on Archive Team’s wiki, and archives it using ArchiveBot (discussed later).

Once the project is selected, core Archive Team members then look into ways to discover and break down the targeted site into manageable items to archive. Describing the overall process, drew quickly simplifies and summarises the infrastructure of an Archive Team project:

“And what we do is we split this large website up in little pieces – so for a social media website that would be each page of the website is one profile, or for a forum each piece of the forum is a single thread. So yea what we have then is a project, and multiple people run the software that I write and they get little pieces of the website assigned to them. Then they download little pieces of the website and send them back to us to our main server. And that main server backs it up into larger files and sends it to Internet Archive.”
(drew)

For the Tumblr project, most blogs have domain names that resemble the following format: *blogname.tumblr.com*, with some users opting to port custom domains to the Tumblr platform. As the primary concern of the project was to capture NSFW content in danger of imminently going offline, Archive Team tasked themselves with figuring out how to source (and select) the blog names to be archived. Multiple strategies were employed. They collected lists of Tumblr domains from various online sources, as well as solicited nominations from the public. Lists of domains were collated in Google Sheets and shared in #tumbledown before they were eventually filtered,

⁴²<http://www.reddit.com/r/shutdown>

⁴³Archive Team have used semi-automated ‘bots’ or agents via IRC and Twitter to facilitate archival nominations for specific sites or media - for example, to allow users to nominate Vine videos for preservation after the announcement that Twitter was shuttering the service (Archive Team, 2016; Vine, 2016).

modified (to suit the format needed by the Tracker) and housed in a central repository in GitHub.^{44,45}

Several seed lists were captured from other scraping efforts, including the pushshift.io Reddit comments and submissions API,⁴⁶ and the majestic.com *Majestic Million* SEO index of the top million domains with ‘the most referring subnets’.⁴⁷ Some seeds were derived from manually collated lists by other community-based curators such as those copied from the Derpibooru *My Little Pony: Friendship is Magic* (MLPFiM) fan community.⁴⁸ The use of these seed lists reveals a couple of aspects about Archive Team’s work and the ramifications for understanding the origins of the archives they produce. It highlights a culture of scavenging and reuse amongst Archive Team, where curated lists from fan communities like MLPFiM, and other collated data sets are repurposed for seeding web archives. However, although it was sometimes possible to initially ‘forensically’ trace the sources of seeds (using GitHub, the Google Sheets notes and Reddit posts) this exercise demonstrated the ways that collection activities are often further obscured with every step towards the act of archiving. Without knowing where to look for the traces of Archive Team’s scavenging practices, once these archives are deposited in the Wayback Machine it becomes impossible to fully understand the various paths by which seeds entered the Archive Team’s collection at the Internet Archive.

This ‘bricolage’ of seed sources reveals some of the ways that Archive Team repurposes other indexing efforts, but also the data work involved in converting these lists into usable datasets for the Tracker to use in the Warrior project. Steps were taken by Archive Team members and participants to collate, de-duplicate and filter millions of links in order to not waste efforts on capturing the same resources multiple times, but also to filter the lists to only include blogs that were tagged as NSFW and ‘adult’. There were however, risks inherent to this approach to seed selection, both for the representative coverage contained within the archives and for the volunteer participants themselves. For example, confusion ensued concerning the difference between ‘adult’ and NSFW Tumblr content, including questions about whether or not the selection of one category would be subsume the other. These discussions (along with the discussion regarding *notes* in the following section) revealed the local, platform-specific knowledge required to disentangle the affordances (and

⁴⁴<https://github.com/ArchiveTeam/tumblr-items> (visited on 9th Mar. 2019)

⁴⁵One collated list provides a record of (in-flux) notes pertaining to the source and number of blogs in each list, confirmation that it should be added to the Tracker and links to the corresponding GitHub repository file. Examining the spreadsheets and associated files allowed me to trace some of the origins of the seed lists, as well as the steps taken to modify the lists to suit the requirements of the Archive Team infrastructure.

⁴⁶<https://pushshift.io/>

⁴⁷<https://majestic.com/reports/majestic-million> (visited on 3rd Mar. 2019)

⁴⁸Since 2015, MLPFiM users have been maintaining a running list of in/active NSFW tumblelogs, as a mechanism for sharing artwork related to MLP fandom. <https://derpibooru.org/forums/uppers/topics/%E2%9C%BF-upload-resources-tumblr-nsfw-active> (visited on 3rd Mar. 2019)

use) of Tumblr in order to archive it. Further concerns were raised by some volunteers about the inadvertent and perhaps unavoidable selection of child pornography as part of this wide-scale approach to nominating NSFW blogs to archive. Throughout the project questions were raised in IRC about Archive Team's position on the legality of such an operation if indeed child pornography was being downloaded using the Warrior application (on to a participant's personal computer). Here, it is notable that after spotting a highly suspect blog (with a characteristically suspicious domain name) being archived by their machine, one participant informed the channel that they were quitting the project altogether.

Further to an initial selection of seeds, there was much debate concerning the boundaries of what should be included and excluded throughout the process of archiving NSFW blogs. The scope of the project was discussed repeatedly on the project channel – discussions which were steeped in an impending (and ever-increasing) sense of urgency, typical of Archive Team's work. One particular discussion regarding blacklisting image-hosting sites external to Tumblr was indicative of concerns that 'hot-linked' images (e.g. images embedded in Tumblr posts hosted elsewhere) did not fall within the remit of this particular project:

```
<frankie> are you manually blacklisting these domains?
<@ezra> frankie: yes I am
<frankie> i.piepec.ru, 1tbporn.net, i.homexvideo.com...
<frankie> I assume a lot of these existed for a grand total of 6 weeks
<frankie> I also got a bunch of errors on 4chan and photobucket hotlinks, but
I forget if they were of the instant skip variety
<grayson> are we supposed to insert content from other domains into the [Internet
Archive]? that feels funny and beyond scope
<frankie> actually yeah why are we pulling that anyway, that's not generally
at elevated risk of deletion
<hayden> a lot of these image upload sites are at risk of disappearing without
warning
<hayden> so if people are embedding images from them I do think it's a good
idea to grab them
<frankie> but *now*, in *this* project?
<@ezra> Well...if it recurses to them why not?
<frankie> anyway does tumblr even let you embed content still? if not roughly
100% of 4chan links should 404
<grayson> sure, but they should be a project in themselves
<grayson> random other sites from years ago an invitation for stalled crawls
<grayson> the crawler only has a single path of execution right now
```

This discussion about blacklisting continued on, as the channel grappled with the most economic ways to code a mechanism for excluding external sites/URLs that

were clogging the crawls and contributing to slow progress. This exchange illustrates just one instance of the dynamic decisions made by those adjusting the crawling scripts about what should be considered ‘in scope’. Further, grayson’s comment alludes to fuzzy/unanswered questions about what (if anything) is agreed with the Internet Archive about the boundaries of the Tumblr NSFW collection activities. Here they are balancing the representative value of grabbing embedded images/resources (even if linked from elsewhere), the perceived likelihood that certain image hosting sites are ephemeral (and therefore likely unavailable) and the effects this has on crawler speeds and completion success rates. In this instance, the observation that the ‘crawler only has a single path of execution’ is a path dependency that forces the team to subsequently prioritise and adjust the scripts to avoid crawler traps in realtime.

Overriding concerns about having enough time to complete the crawls before Tumblr’s removal of NSFW sites were ever-present in all decisions made about adjusting the crawl parameters of the Warrior scripts. Although the above example about black-listing sites implies an importance placed on archiving platform components deemed important by its users (e.g. ‘if people are embedding images [...] I do think it’s a good idea to grab them’), other discussions about archiving different components of tumblelogs (e.g. *notes*) further revealed the ways that Archive Team collectively evaluates and forms ad-hoc consensus around other types of exclusions. The example of Tumblr ‘notes’ presents another salient example of this form of valuation. Helpfully, Seko and Lewis (2018) explains the role of notes on the Tumblr platform:

“While giving another post a ‘like’ is a gesture of affirmation prevalent among social media platforms, the ‘reblog’ feature contributes largely to Tumblr’s unique media ecosystem. By reblogging a post, bloggers can copy and repost the material made by others on their own dashboards (i.e. homepages). The record of interactions with a post is immediately attached to the post through ‘notes’ that list the original poster and each user who has reblogged or liked the post” (Seko and Lewis, 2018, p.183).

Here, in order to decide on whether or not to archive notes, participants discussed the form and function of notes on Tumblr – which given Tumblr’s unique system of re-blogging, was not readily apparent to all involved in archiving the platform. This was then followed by many discussions on whether or not notes were a priority to archive, with most deliberations hinging on the lack of time available to crawl. Over the course of several days, various participants made the case for different views on the matter – to continue archiving notes in full (despite the time constraints), to partially archive notes (e.g. only the first 50 on any post) or to fully exclude them from the archive. Ultimately, in a move to speed up the captures and in the absence of definitive technical paths for efficiently including a partial capture of notes, several participants announced what they deemed ‘The DecisionTM’ to exclude notes from the crawls:

5.3. Archive Team at Work

<ezra> I am all for ignoring notes at this point, since we have so little time and some sacrifices must be made to get as much as we can

[...] (*two days later*) [...]

<indiana> most Tumblr file metadata consists of "RandomDerp liked this" and "RandomDerp reblogged this image". It is 99% worthless.

<jamie> indiana, I'm actually interested in the who reblogged it stuff. that's a snapshot of the community... the social network... and that's a huge part of what they're destroying here.

<kyle> indiana: well that's just a fundamental disagreement then. I don't think preserving just a part of history is preservation at all.

<kyle> agree with jamie

<logan> indiana: Destroying the like and reblog metadata destroys the social part of the social network that is Tumblr.

Here ezra characterises the exclusion of notes as a necessary 'sacrifice' given the implications of time and resources to archive Tumblr. Several days later participants were still discussing the value of notes as a proxy for understanding the social dynamics of the Tumblr NSFW blog network. Here members are balancing the desire to collect more (a trait I call 'archival abundance', discussed further in the next section) with the value of collecting particular components of Tumblr blogs. The continued discussion reveals the ways that Archive Team participants both contest decisions and work to frame the value of collection through the lens of a desire to produce a complete and representative record of 'the social network'. This desire is also framed through opposition to the actions of Tumblr to remove the NSFW posts in the first place, an act of 'destruction' that comes at the detriment of a future understanding of this community if seen through the use of web archives that only partially captured the experience of Tumblr.

Furthermore, the issue of notes is reflective of concerns over the ways that slow crawling had an effect on Archive Team's ability to recruit new participants and ultimately, collect more Tumblr blogs:

<austin> I will not scream louder until I know we have the warriors moving faster

<grayson> I agree with austin. its [frustrating] to see slow crawls

<grayson> people might not come back

These examples also work to illustrate concerns from Archive Team that slow crawling could potentially turn away possible volunteers (who would rather see the Tracker speeding along) and therefore limit participation in the project. grayson reflected that it was 'frustrating to see slow crawls' and a further concern that 'people might not come back'. Over the course of the Tumblr grab, participants periodically tested the limits of the targeted infrastructure, in terms of its bandwidth capacity and the

potential likelihood for rate-limiting for particular user agents if used at scale. The slow decay of many sites targeted by Archive Team (neglected by time and infrastructural resources) often makes them incapable of supporting the bandwidth required to enable simultaneous large-scale access.

In the case of Tumblr, however, it would not be the slow decay of their server infrastructure that would impact the rate of archiving, but rather, what Archive Team interpreted as multiple concerted efforts (on the part of Tumblr itself) to derail and block their efforts to archive NSFW blogs before they were taken offline on the 17th of December. Tumblr's efforts involved rate-limiting, but also user-agent throttling and wholesale IP bans that led Archive Team and volunteers to continuously investigate, test and implement work-around solutions in order to continue crawling NSFW sites, particularly over the course of the last few days before Tumblr retracted public access to NSFW blogs. This game of 'cat and mouse' and the strategies undertaken are further discussed below.

Circumventing the Ban, Brute Force Archiving in Practice

In the case of Tumblr, a breakdown in crawling became a moment when internal assumptions about practices were revealed, discussed and subsequently re-configured based on community conventions and the realtime collective priorities of those doing the work. Over the course of the two weeks, Archive Team participants were repeatedly rate-limited and/or banned by Tumblr. As discussed earlier in this chapter (Section 5.2.2) it is a regular occurrence for Archive Team to receive pushback from the sites being targeted by their crawling efforts. As drew explains:

"Yea we are banned sometimes. We have some nice examples of the owner of a website coming into our IRC channel – and then we use the user-agent 'Archive Team' so you know, it's not impossible to know who is causing such a high load on their servers. So sometimes people come into our channels and ask what we are doing, and if they can help us, for example. But sometimes people are a little aggressive and they decide to ban us. And what we usually do is throw more IP addresses at them, use more common user-agents so it's a little harder to automatically ban." (drew)

As explained by drew, Archive Team typically uses a bot that is identified through the 'archiveteam' user-agent which makes web masters aware of the origins of increased access requests to their servers. For the Tumblr project, breakdowns in crawling caused by mass IP bans and significant rate-limiting of various user-agents, forced repeated discussions around whether or not Archive Team permitted the use of logins and different (more common) user-agents during crawling. Here logins or the use of additional cookies that enabled crawler bots to mimic 'real users' became a topic of extensive debate that revealed an ethical dilemma for Archive Team members who attempted to balance a desire for abundance (e.g. more crawling, more archives)

with ensuring particular forms of archival integrity. For newcomers like grayson, it became unclear whether or not the question and use of login cookies in Archive Team was in fact, what they distinguished as a *technical* issue or one of *policy*:

```
<grayson> drew said there was some decision about login cookies but didn't get
to explain the why

<ezra> I would also like to know the reason why we can't do login cookies

<grayson> we had at peak 4 people working on code changes for it so we need
to repurpose them to something we're going to use

[...]

<austin> Using login cookies apparently crosses a line

<austin> Then we're acting like people

<austin> And it gets into the WARCs

<mica> Hm, does that mean that there's no way to get around the ratelimiting
after all?
```

Only through further questioning ezra (after the project had finished) was I able to understand the reasons underpinning the debate about logins. Here, the discussion of logins reflects a general Archive Team 'policy' that I have not seen documented elsewhere:

```
<ezra> With logins/cookies we are sort of contaminating the resulting warc
with data that is not really meant to be there. I'll give an example, if Facebook
decided to shut down their service, since posts and profiles are quite easily
viewable publicly without authentication we can grab that, however with private
profiles we risk grabbing data that is not

<ezra> meant for public consumption, such as a users email address, or an users
mobile phone number(that would be disastrous)

[...]

<ezra> However if a site was previously publicly viewable, and they now have
a login wall, we may circumvent that by using a login, but we don't like doing
it as we are linking an account to what is supposed to be anonymous data

<ezra> By just using a login we expose data that shouldn't otherwise be exposed,
ie in my example user email addresses, mobile phone, secret questions, etc
```

Here ezra outlines the reasons why they tend to steer clear of using logins to crawl, namely to avoid the risk of collecting components of websites that are typically only viewable through the use of access controls. Through by-passing these access controls, Archive Team risks inserting information into the WARCs that was once only viewable behind a login, through friend networks and cookies that enable both. But in the case of Tumblr, and in the face of the prospect of having to stop the project and the collection of NSFW blogs, participants ultimately decided to use login cookies – in order to achieve this an entire infrastructure was built to support the dynamic generation of cookies (called 'a cookie factory'). The choice to proceed with the use of login cookies reveals a hierarchy of priorities for Archive Team participants, and

therefore two observations about community practice and culture. First, despite a general ‘policy’ against the use of logins, the prioritisation of continued crawling over abandoning the project reveals a fetishisation of abundance that ultimately trumps the risks imposed by breaching access restrictions. Here, the desire to collect more – whether in terms of scale or completeness – was reflected in many aspects of Archive Team’s practices, including the use of the Tracker leaderboard and the Warrior categories of success.

Second, the juxtaposition of these two priorities – what we might call *archival abundance* and *archival integrity* – reveals the ethical boundaries and practice values of the community that both relies on conventions of practice and core members to translate them to new participants, but is also indicative of a community that promotes *adaptability* (Kelty, 2008) in practice. Here adaptability is associated with ‘the ability to critique design and propose alternatives’ (Kelty, 2008, p.235) that reflects the dynamic and performative ways that web archiving is both shaped by culture (through the symbolic and material conventions and strategies of action) but also works to transform community practice, and as a result, the archived Web they produce.

5.4 Chapter Summary

This chapter offers significant insights into the research question: ***In what ways do web archival practices (the who, why and how) shape the archived Web?*** Through the lens of *web archiving as culture* two observations about web archiving were made. The first is that web archives, like culture, are enacted through practice. Through the case of Archive Team, I have shown that web archiving is produced within particular cultural worlds, systems of meaning and strategies of action. Second, web archiving is thereby transformative, in that it both creates and sustains a community of practice but also inevitably shapes how and why the Web is archived.

In the first half of the chapter I framed the cultural dimension of Archive Team’s web archiving practices through the concept of community. I introduced Archive Team through a discussion of their origins in the closure of AOL Hometown, and the subsequent ways Jason Scott and others formed a community of practice in pursuit of archiving the Web. The use of IRC and the wiki to organise Archive Team activities not only stems from a pragmatic need for devices that enable distributed collaboration online, but are also indicative of underlying epistemological commitments to decentralised ways of working that reflect a desire to self-document and promote volunteer contributions to their collective goals. I argued that community protocols around organisational self-archiving and IRC communication were indicative of practices that communicated and enrolled new participants into a culture based on two tenets of what I’ve called ‘radical web archiving’. Here I proposed the concepts of *archival neutrality* and *brute force archiving* as two tenets that shape the ways Archive

Team frames their own practices, particularly in opposition to more conservative approaches to web archiving.

The second half of the chapter detailed Archive Team's web archiving practices through the context of their efforts to archive Tumblr NSFW posts in December 2018. Here, I contextualised their efforts through a longer history of Archive Team's engagement with Yahoo, as well as a detailed summary of the sociotechnical aspects of the Tumblr platform and their policies towards NSFW content. I described the use of the Warrior application and efforts by Archive Team to mobilise a wider public to the cause of web archiving Tumblr NSFW blogs. The final section returns to the concepts of 'neutral' and brute force archiving as a mechanism for exploring how these tenets of the Archive Team approach both shaped the nature of the archived Tumblr blogs, but also revealed the ways that practice is negotiated, contended and enacted through the circumstances and challenges of web archiving. These final sections revealed a hierarchy of priorities, including a quest for balancing the desire for what I called *archival abundance* and *archival integrity*. These priorities, as observed through the decisions surrounding whether or not to archive particular components of the Tumblr platform ('notes' and hyperlinked images) and use login cookies in response to attempts to ban the group by Tumblr, revealed a community actively engaged with the ramifications of their practices. Participants regularly discussed and contested aspects of their collection practices, whilst overwhelmingly favouring approaches that enabled them to collect more, despite the risks of breaching the access protocols of sites like Tumblr.

In conclusion, this chapter has revealed the ways that Archive Team is actively shaping the nature of what they are collecting, as well as creating a cultural world and community of practice that is steeped in a tradition of 'radical' approaches to archiving the Web. Despite the use of specific standards of web archiving like the WARC format, these practices will be seen in stark contrast to other approaches taken by conventional memory institutions. However, given the scale of the impact of Archive Team's collecting activities on the shape of the Internet Archive Wayback Machine, understanding their practices enables crucial insights into the ways that the Web is transformed through web archiving.

The next chapter, Chapter 6 extends the study of community web archiving to examine *web archiving as politics* in the context of the Environmental Data & Governance Initiative.

“There is no political power without control of the archive, if not of memory. Effective democratization can always be measured by this essential criterion: the participation in and access to the archive, its constitution, and its interpretation.”

JACQUES DERRIDA (1998, *Archive Fever*)

6

Web Archiving as Politics: Environmental Data & Governance Initiative

6.1 Introduction

This chapter explores *web archiving as politics* through the case of the Environmental Data & Governance Initiative. In November 2016, in the wake of the election of Donald Trump for President of the United States, an international grassroots network of environmental specialists, lawyers, librarians, archivists, civic technology advocates and concerned citizens mobilised to archive data and web pages pertaining to US federal government environmental policy, climate science and research. Continuing through June 2017, these ‘guerrilla archiving’ events (later dubbed the ‘*DataRescue* movement’) formed part of a strategy by the newly formed Environmental Data & Governance Initiative (EDGI) and collaborators, to centre web archiving as a key mechanism for mitigating the potential effects of an incoming US administration who were seen to be ‘anti-science’ and ‘anti-environment’.

The chapter uses the lens of *web archiving as politics* to emphasise some of the ways that web archiving – like other forms of memory work – is inherently a political project. This point has been made elsewhere through decades of engagement within the archival profession with the tensions between the view (by some) that recordkeeping is and should remain an objective and fundamentally neutral endeavour (Greene, 2013), versus those that have challenged these claims, to recognise the power and politics embedded within the processes of selecting and preserving the past for the

future (Brown and Davis-Brown, 1998; Caswell, 2009, 2013; Cook and Schwartz, 2002; Zinn, 1977). Web archiving as politics is premised on Harris's argument that 'the archive is politics – not that it is political, but that it *is* politics' (2005, p.173, emphasis my own). This claim is distinctive from other assertions that acknowledge the 'political nature of archives', and the ways that archivists are actively implicated in the political processes of archiving and recordmaking as a form of power (Harris, 2005, pp.174-175). 'Archives as politics' builds on the work of Derrida (1998) and Foucault (1972) to assert that political power fundamentally stems from the control of information and memory – where power is exercised through the construction and control of 'the archives'. Politics is therefore a lens to focus the discussion on the conditions of possibility that web archiving creates, where the construction of the archive is negotiated, contested and ultimately becomes the 'law of what can be said' (Foucault, 1972, p.145).

Building on this, I make two over-arching observations about *web archiving as politics* that further emphasises the mutually constitutive ways that practices (who, why and how) shape the archived Web. The first is that politics work to motivate and mobilise web archiving as a tool for both ensuring access to the Web, but also as a mechanism for enacting particular claims about the ways the Web should work. This observation is extended by a second point which supports the notion that politics also emerge through the practice of web archiving and the everyday negotiations and decision-making that determines how and what of the Web is saved.

Politics of Practice/Practice as Politics

The chapter begins by examining politics through a discussion of the emergence of EDGI and the ways they framed the perceived incoming threat posed by the Trump administration to environmental and climate data, in particular. Narrated through a *crisis* discourse, I contextualise the origins of EDGI within a Political climate of uncertainty that motivated both the formation of the organisation and a particular form of 'resistance' that emphasised the urgency and vulnerability of environmental policy and federal public data infrastructures. 'Crisis' provides a motivation and rationale for EDGI's work, playing both a structural role in the chapter and an analytical opportunity for understanding web archiving as a set of situated practices that are deployed in response to specific geo-political circumstances – in this case, the US presidential election.

Second, through the DataRescue movement, web archiving becomes a boundary object for mobilising a particular form of politics that is enabled through the act of 'rescuing data'. In Section 6.3, I use popular STS concepts of *boundary objects* (Star and Griesemer, 2015) and *boundary work* (Gieryn, 1983) to frame the ways the DataRescue movement used web archiving as a means to bring together different groups around a common goal. Here, fears surrounding the potential disappearance

of web-based environmental data initially mobilised the involvement of over 1,500 participants (Lamdan, 2018) in volunteer efforts. EDGI and collaborators worked to design and develop interventions that translated the needs of multiple forms of expertise to archive sites and data. The politics of expertise are revealed through boundary work, or the ways that web archiving is negotiated by a diverse set of sociotechnical actors from a range of disciplines, professions and civic interests.

And in the case of EDGI, the US presidential election and political party Politics – what I’m calling ‘big P politics’ – is certainly implicated in the ways that web archiving was mobilised in DataRescue. But ‘politics’ is also used to capture the everyday ways that ephemerality, expertise and resistance are all negotiated, contested and manifested through web archiving practices. In this case, web preservation can be seen as part of a wider geo-political climate that draws attention to the power relations underlying the production of evidence, scientific facts and truth-making. In the case of EDGI, web archiving has been deployed as a mitigation tactic for both the loss of public data essential for researching environmental harms and climate science, but also as a means for critically engaging with the role of the state in the production and preservation of science data. Here web archiving enables alternative forms of activist interventions that move beyond institution-centred approaches to web archiving. Many issues were encountered by DataRescue participants and organisers, including the sociotechnical challenges associated with the (in)efficiencies of crawling technologies, the verification of archived data and concerns over selection (Lamdan, 2018) – all of which impacted how web archiving was negotiated and enacted as part of the DataRescue movement. But in this context, web archiving and EDGI’s associated projects also worked to empower communities in the creation of alternative imaginaries and infrastructures for data stewardship and relations (Dillon et al., 2019). Below, this chapter begins by framing the emergence of EDGI in the wake of the US Presidential election in 2016.

6.2 Origins of EDGI, Framing a Crisis

In their own words, the Environmental Data & Governance Initiative is ‘an international network of academics and non-profits addressing potential threats to federal environmental and energy policy, and to the scientific research infrastructure built to investigate, inform, and enforce them’.¹ Since its formation in November 2016, EDGI has been working to mobilise citizens in both the United States and Canada around collective action dedicated to protecting and critically engaging with environmental data governance. To further these aims, EDGI members have interviewed former and existing US federal employees, built archiving and analytical tools for monitoring US

¹<https://enviropdatagov.org/about> (visited on 15th Dec. 2017)

federal web pages, tracked legislative changes and submitted Freedom of Information Act (FOIA) requests, and organised events and ‘research networks to proactively archive public environmental data and ensure its continued public availability’.² Since 2016, EDGI has evolved to include several working groups and projects focused on Interviewing, Archiving and Community Technology, Environmental Data Justice, Capacity and Governance, and Website Monitoring activities. In addition, members rotate on and off of a Steering Committee which oversees the continuity of the organisation. Each group has its own distinct working rhythms, activities, roles and projects however, each is ordered by EDGI’s overarching principles of horizontal and consensus-based governance subscribed to by the 175 scholars, activists and civic technology advocates which make up EDGI’s membership.³

This section is fundamentally led by a desire to unpick the ways that the contemporary political climate of US President Donald Trump’s ‘anti-science’ agenda sparked a social movement around what was framed as a potential crisis for environmental protections and research in the US. Over the course of this research, participants indicated that a particular mode of working came to characterise the first year of EDGI’s organisational existence. In the context of discussing the work of web archiving, one EDGI member, Bernard reflected that ‘in a project like EDGI there’s a lot of urgency, we work in crisis mode’. As a mechanism for discussing the political dimensions of web archiving, this section reflects on the emergence of EDGI and their initial web archiving projects through a *crisis* discourse. Here I use ‘discourse’ to refer to the ways that ‘statements say something about individual or socially shared *subjective reality*’ – rather than as a mechanism for producing a particular kind of effect or outcome (Alvesson and Kärreman, 2000) – where ‘crisis’ is indicative of the ways that participants (as social actors) reflexively represented practices upon reflection (Fairclough, 2012, p.455). Discourse is therefore used to make a connection between the ways that participants collectively framed the election of Donald Trump as a moment of crisis (both politically and environmentally) and the implications this had for why and how web archiving was mobilised to further EDGI’s aims. For context, I draw on Rosenthal, Charles and ‘t Hart (1989), who define a crisis as:

“[...] a serious threat to the basic structures or the fundamental values and norms of a system, which under time pressure and highly uncertain circumstances necessitates making vital decisions” (Rosenthal, Charles and ‘t Hart, 1989).

Using three elements of crises outlined by Boin and ‘t Hart (2007, p.43-44), this section explores the *threat*, *uncertainty* and *urgency* that motivated the centring and use of web archiving by the newly formed EDGI organisation. As will be examined, EDGI’s efforts were rooted in both a threat to federal environmental science data posed by

²<https://github.com/edgi-govdata-archiving/overview> (visited on 15th Dec. 2017)

³As of a 2018 EDGI scholarly publication, the organisation reported it had 175 members (Vera et al., 2018, p.511).

the Trump campaign's public denouncements of climate change (Dennis, 2016b), but were also tied to historical precedents for dismantling environmental protections set by previous US and Canadian governments. Below, participants describe EDGI's early engagement with web archiving, and the ways these responses were framed by a sense of urgency and uncertainty that were manifested in the material and emotional labour of 'empowering a broad community to work together to copy and preserve data that they cared about' (Dillon et al., 2019, p.7). Further, by acknowledging the subjective processes behind the identification of crises (Boin and 't Hart, 2007, p.45), this framing foreshadows emergent tensions between EDGI's goals of 'rescuing data' and the mixed reception of these activities as 'an extreme reaction' by some in the environmental science and wider web archiving communities (Brennan, 2016). Here, politics shift between the centring of web archiving as a tool for ensuring data 'accessibility amidst times of political instability' (Walker, 2017a), to the recognition that web archiving practices themselves are situated, contested and subject to the expertise and political will of a coalition of stakeholders that shape *why* and *how* web archiving is done.

6.2.1 The Threat of Donald Trump and the Harper Years

"Crises occur when the core values or life-sustaining systems of a community come under **threat**. [...] The more lives are governed by the value(s) under threat, the deeper the crisis goes (Boin and 't Hart, 2007, p.43, emphasis my own)."

In the weeks following the 2016 US Presidential election, EDGI emerged through a collective email exchange amongst a small network of humanities and social science scholars, and environmental organisations (Vera et al., 2018). The founders of EDGI organised around a concern over 'the future of environmental science, data, and policy in the face of a virulently anti-science and anti-environment administration' (Dillon et al., 2019). For years prior and throughout the Presidential campaign, Donald Trump publicly denounced climate change science, claiming that he was 'not a believer' (Dennis, 2016b) and appointing vocal deniers of climate change to the transition team who openly vowed to defund the Environmental Protection Agency (EPA) (Davidson, 2017). During the period immediately surrounding the Presidential inauguration, EDGI conducted extensive interviews which exposed a demonstrable fear amongst current and former career staff at the EPA and the Occupational Safety and Health Administration (OSHA) that the Trump administration 'posed the greatest threat to the [EPA] in its entire 47-year history' (Sellers et al., 2017). Moreover, EDGI's framing of the threat posed by the incoming administration (and therefore, the nature of their response) was informed by their own collective values, expertise and knowledge about historical precedents for governmental tactics for dismantling

environmental protections. One EDGI academic and early collaborator, Claire, describes their experience of framing the politics of information access in a ‘very intense discussion’ with their students, specifically around whether or not it was ‘game over on climate change’. They went on to describe their motivations for the Web Monitoring project as emerging based on (and explicitly grounded in) their fears surrounding the potential disappearance of government web pages that they relied on for teaching:

“[...] the class was on mitigation via switching out energy sources and the exercise was designed all around information from the Energy Information Administration which is a US website that’s part of [the Department of Energy]. And it’s supposed to be a source for impartial information on the environment. And so I had students do some work before they came into class and they were working through data collectively from that site during class. And as I was walking home from class [...] I was thinking, wow when I teach this class next semester I’m not sure if I’m going to be able to use that website.” (Claire)

Beyond a reliance on particular websites for teaching, this comment is reflective of an uncertainty surrounding the potential effects of the Trump administration on the impartiality of environmental information supplied by federal agencies. Whereas others have characterised the removal of web content related to climate change as ‘expected’ on sites like whitehouse.gov (Brügger, 2018, p.2), Claire points towards a fundamental difference between the removal of information related to the political priorities of a particular administration and the removal or denial of scientific facts by federal agencies charged with providing impartial information access to the public. Reactions to the administration are further contextualised in recollections by Bernard, who again recalls a collective ‘sense of serious crisis’ that immediately followed the election:

“So you know, the election happened and it was quite a system-shock for a lot of people, certainly for me also. [...] It was a feeling kind of reminiscent of when George Bush was re-elected in 2004. Like what just happened? You know? Who could make these decisions? So I think that was a common feeling of like whoa, body blow—and what does this mean for the things that we care about for a lot of the people involved in EDGI. And that, you know, that’s really pivotal to the success of DataRescue. That sense of—in that first 3 or 4 months after the election—of serious crisis before the Twitter presidency was normalised. You know, a kind of disbelief that the United States would be governed in the way that in fact, it’s being governed now.” (Bernard)

Here Bernard acknowledges the unexpected nature of the election results, through

a description of their visceral response ('system shock', 'body blow') to an imminent threat directed at 'the things we care about'. Bernard's observation that they previously felt this sense of crisis works to highlight other precedents that participants observed in their reactions to the threat posed. As EDGI has documented elsewhere (Sellers et al., 2017), Canadian Prime Minister Stephen Harper's administration (2006 - 2015) provided a recent reminder of the dangers of canceling climate change programmes, 'muzzling' federal environmental scientists and widely removing public access to information through federal websites, libraries and archives.⁴ With early EDGI members spanning between Canada and the US, participants acknowledged the impact of the Canadian precedent on their own sense of the threat:

"I had this Canadian precedent in mind. Of how quietly and quickly, and in my mind, effective to their goals [the Harper] government was in strategically shifting the whole bedrock around what information was available to people about environmental things. But also they abolished the long-form census—census data which is used by so many different disciplines and used as baselines against which tons of community organisations and NGOs [non-governmental organisations] and other people make policy recommendations. And so I felt a sense of concern because [...] I don't remember it being so public of an issue in terms of [Stephen] Harper's [administration]. He just never said stuff in the same way. [...] Whereas with Trump it was so vocal. These are targets. It felt like such an active targeting." (Augustine)

"Yea so in Canada there wasn't a census one year. And the term libricide was invented in Canada. And that happened in my time, I remember those years and I remember how shitty that was. So when this stuff came along I was like this is real, this will happen and to me it was just a foregone conclusion that they would start censoring data." (Luke)

Here participants consider the ways that the precedent set by the Harper administration framed their own responses to the threat of Trump, with Augustine comparing the two administrations and reflecting on the public sense of 'active targeting' that Trump, in particular, conveyed with regards to the environment. The removal of the mandatory long-form census under Harper⁵ is raised by both participants – pointing towards the ways that members were framing the significant *role of data* (and its availability) in the documentation of environmental harms and the formation of national agendas for environmental protections and policy. Luke expresses a fear surrounding the threat of 'libricide' – a term used to signify the 'regime-sponsored,

⁴See Coates (2015) for an entire issue of *Canada Watch* in response to the Harper government's suppression of the study and dissemination of knowledge about Canada.

⁵Under the Harper administration, the 2011 long-form census was replaced with a short-form census and a longer voluntary survey (known as the *National Household Survey*). Over 500 Canadian organisations protested the decision, including the Canadian Medical Association and the Canadian Chamber of Commerce (Marche, 2015), where the removal was argued to be a tactic for disabling the collection of data (particularly on socially/economically marginalised populations) that could be used to support social advocacy causes in government (Murdoch, 2010).

ideologically driven destruction of books and libraries' (Knuth, 2003, p.5). Here Luke indicates a concern for the *role of libraries and archives* in the stewardship of cultural and scientific memory, but also acknowledges the precariousness of these institutions (and their holdings) when confronted with political ideologies that centre the availability of information as a tactic for political gain. Although EDGI members were aware of and in contact with libraries and archives involved in existing efforts to archive the .gov webspace (discussed below) there remained a certain degree of uncertainty over the coverage of existing institutional programmes (particularly around datasets), as well as how communities would be able to access these archives in the future.

The next section reflects on how these uncertainties, coupled with the urgency of the political climate led EDGI and other emergent organisational and community-based projects to mobilise their own collaborative efforts to use web archiving as a mechanism for responding to the incoming administration.

6.2.2 Uncertainty, Urgency and the End of Term

"In a crisis, the perception of threat is accompanied by a high degree of **uncertainty**. This uncertainty pertains both to the nature and the potential consequences of the threat: What is happening and how did it happen? [...] What can we do? What happens if we select this option? What will others do? (Boin and 't Hart, 2007, p.44, emphasis my own)."

"Crises induce a sense of **urgency**. [...] Time compression is a defining element of crisis: the threat is here, it is real, and it must be dealt with as soon as possible (Boin and 't Hart, 2007, p.44, emphasis my own)."

Many EDGI participants described the presence of uncertainty and urgency during the first few months of the organisation's existence, tied to the perceived threats to federal web-based environmental data and information access. Augustine reflected that before there were EDGI working groups, or regular weekly (virtual) meetings, there was an everyday climate of urgency that persisted during the months before Donald Trump's inauguration:

"[...] *that everyday was pretty—it just felt so urgent. That's the other thing. It's pre-inauguration, we don't know what's going to happen. And it felt pretty loose. I feel like there was a less strong understanding of how or what identifying as an EDGI person exactly meant.*" (Augustine)

Alongside EDGI's other early initiatives (particularly around interviewing federal workers) the work of imagining and coordinating an events-based model for web archiving became an organising mechanism for both the articulation of EDGI membership and resistance of the impending uncertainty that surrounded the transition period. In November 2016, EDGI founders and early members identified a potential collaborative partner in the *End of Term* web archiving project. Running since 2008, the End of

Term (EoT) project is an ongoing cooperation between several institutions that collectively seed and crawl the US federal domain during Presidential transition years.⁶ The University of North Texas's nomination tool provides an interface through which the public can play a part in the nomination process,⁷ which enables a 'community of subject specialists with a convenient means to contribute information on specific sites for the focused crawl'.⁸ Accordingly, the tool allows users to submit particular URLs (or *seeds*) for crawling by the EoT infrastructure, whilst providing basic information about the URL (relevant branch of government and agency) and themselves (name, institutional affiliation and email address) in the process.

In previous years, the EoT crawl received relatively little public engagement through the nomination tool, with the project reporting around 30 nominators in both 2008⁹ and 2012.¹⁰ As several EDGI members have noted in interviews, EDGI was not the only organisation which formed around a perceived threat to environmental data in the wake of Trump's election. Here Augustine attributes part of the emergence of these independent projects as a response to an uncertainty surrounding what would be covered as part of existing web archiving efforts like the EoT crawls:

"[...] we were already interacting and engaging with Internet Archive and trying to think about a correct approach, but there was so little that was clear about any assertion about how much would be archived. And so that's also why we were not alone, there were those other projects like Project Azimuth and Climate Mirror and Data Refuge that independently were also like 'oh shit, this could be affected'." (Augustine)

Part of this uncertainty can be attributed to the scale of the problem, as well as the complex entanglement of various statutes and mandates surrounding who is in fact responsible for ensuring the preservation of the US federal webspace. EoT project members have spoken publicly about the impossible scale of the task of archiving the government Web in its entirety, at times describing the proliferation of government sites as akin to 'invasive species' (Jacobs and Bailey, 2017).¹¹ Furthermore, as Phillips and Phillips (2019) describe, the US National Archives and Records Administration

⁶The 2016 End of Term project involved the Library of Congress, California Digital Library, University of North Texas Libraries, Internet Archive, George Washington University Libraries, Stanford University Libraries and the U.S. Government Publishing Office.

⁷<http://web.archive.org/web/20170107092748/http://digital2.library.unt.edu/nomination/eth2016/add/> (visited on 22nd May 2019, archived on 7th Jan. 2017)

⁸<https://digital2.library.unt.edu/nomination/eth2016/about> (visited on 9th May 2019)

⁹<https://digital2.library.unt.edu/nomination/eth2008/about> (visited on 22nd May 2019)

¹⁰<https://digital2.library.unt.edu/nomination/eth2012/about> (visited on 22nd May 2019)

¹¹Adding to the issue of scale, the US legal mandates surrounding the preservation of the federal webspace, particularly in times of administrative transition are far from clear. The US Federal Records Act (FRA) which mandates the maintenance, preservation and (where applicable) public notification of the destruction of government records is, as Lamdan (2018) argues, ill-equipped and in need of updating for the regulation and enforcement of electronic records management. Whilst the US National Archives and Records Administration (NARA) is federally tasked with the preservation of government digital records, Lamdan (2018, p.241 n.40-41) notes that their (discretionary) digital archiving strategy 'does not carry the force of the law' and is insufficient in reaching a 'policy consensus on whether all types of webpage materials' are indeed 'federal records'.

(NARA), the Library of Congress and the US Government Publishing Office include ‘web archiving as part of their imperative’ but none of these organisations have a broad enough mandate to capture the whole federal webspace. For this reason, the EoT project formed in 2008, following a memo by NARA that (despite previous attempts) they would not be archiving the wider federal domain during the Presidential transition following the election that year (Phillips and Phillips, 2019).¹² Therefore, in the absence of a funded mandate and despite the scale of the task, the EoT project operates on an ad-hoc, volunteer basis where institutions donate staff time and expertise to the task of archiving the federal websites.

It is within this context that EDGI identified the EoT project as a means for contributing to existing web archiving efforts. But, as Augustine acknowledges, there was still uncertainty surrounding the nature of EoT project and its coverage:

“I find understanding what is in the End of Term collection not clear. So if you were concerned about a specific dataset, which I personally was less so, but there were so many people in EDGI who were because of their research interests and environmental justice background—it was not so easy to have an assurance that this thing that you cared about was there, or it wasn’t apparent to us how to do it.” (Augustine)

Augustine expresses an uncertainty amongst ‘many people in EDGI’ surrounding the fate of digital resources that were central to their research interests, which broadly spans the humanities, social sciences and law, physical and life sciences and ‘justice-centred approaches to environmental data and policy’ (Dillon et al., 2019). For EDGI members, uncertainties surrounding the vulnerability of datasets – particularly related to climate change and environmental health and hazards – manifested both in terms of fears associated with political precariousness of access but also the technical aspects of the crawling technologies used to archive web-based resources. Distinctions emerged between ‘datasets’ and ‘web pages’ that fostered further concerns about what would actually be captured by the EoT project crawlers and how researchers would be able to access the resources of concern (or ‘the thing you cared about’).

These uncertainties with regards to what was being archived (where and by whom) would continue throughout EDGI’s DataRescue work and in their Web Monitoring project. The sociotechnical complexity of the task, combined with the urgency of the political climate and the fears surrounding the impending removal of access to public environmental data and web resources had several effects on the ways that EDGI

¹²In fact, NARA conducted the first large-scale capture of the federal webspace after George W. Bush’s first term in 2004, however the 2008 memo announced their decision to discontinue their practice of snapshotting the wider federal government domain (with the exception of the Congressional and Whitehouse websites which are governed by the FRA and Presidential Records Act, respectively). They argued that NARA’s creation of snapshots would discourage federal agencies from managing the preservation of their own web domains (as NARA declared this the statutory responsibility of the agencies themselves), and further added that they remained unconvinced of the ‘permanent archival value of a Federal agency web snapshot taken on one random day near the end of a Presidential term’ (National Archives and Records Administration, 2008).

manifested politics through web archiving. Whilst the threat of data deletion (and other known tactics for discontinuing access)¹³ worked to foster widespread goodwill and drive EDGI's collaborative interventions in web monitoring and event-based web archiving, it also contributed to the burnout experienced by some volunteer members who – in addition to their day jobs – worked for months in crisis mode. The collective sense of urgency invigorated a DataRescue movement that in a matter of weeks mobilised thousands of volunteers to archive federal data and websites, whilst seeding the work of building inclusive, alternative communities and infrastructures for stewarding public data (Dillon et al., 2019). The next section explores the ways the workflows and tools developed by EDGI and others enabled DataRescue, with a view towards understanding the political dimensions of how web archiving worked to bring together people and technologies to archive federal websites and data.

6.3 DataRescue and the Boundaries of Web Archiving

In December 2016, the University of Toronto's Technoscience Research Unit (TRU) and a fledgling EDGI organisation coordinated the first 'guerrilla archiving' event in Toronto. The event was aimed at several activities: to identify potentially endangered environmental programmes, data and databases; nominate seeds to the End of Term crawlers; establish difficult to crawl data sources and create scripts to archive them; and begin constructing a 'toolkit' for other organisations to host similar future events (Murphy, 2016). Whilst Toronto was the first, 48 DataRescue events followed in cities and universities across the US, including Philadelphia, Indianapolis, Los Angeles, Ann Arbor, Boston and New York City, to name only a few (see Figure 6.1 for the geographic distribution of DataRescue). The public events model worked to bring together an estimated 1,500 participants (Lamdan, 2018) around the cause of DataRescue, spawning new partnerships between EDGI and other science activists, civic technology groups and organisations with data curation and preservation expertise. In this section I explore *web archiving as politics* through two widely-deployed concepts in Science and Technology Studies (STS): *boundary objects* and *boundary work*.

In some ways the following could be read as a contradictory approach to the notion of 'boundaries' in STS. For Star and Griesemer (2015, p.176), *boundary objects* are analytic concepts used to describe 'robust' scientific objects that 'inhabit several intersecting social worlds' and take on a fungibility that enables both a common identity and local adaptation by collaborators. As such, they exist at the intersections, rather than the margins of different social worlds, temporarily bridging heterogeneous practices

¹³In an interview prior to the 'guerrilla archiving' event, one member described the ways that EDGI was not only concerned about the outright removal or deletion of data, but also recognised the broader set of tactics that can prevent access to public environmental data: "If programs aren't cancelled, I expect to see them starved of budget, of personnel, of resources. You can stop data collection that way. I expect datasets might be pulled back out of public access and be harder to get" (Kupferman, 2016).

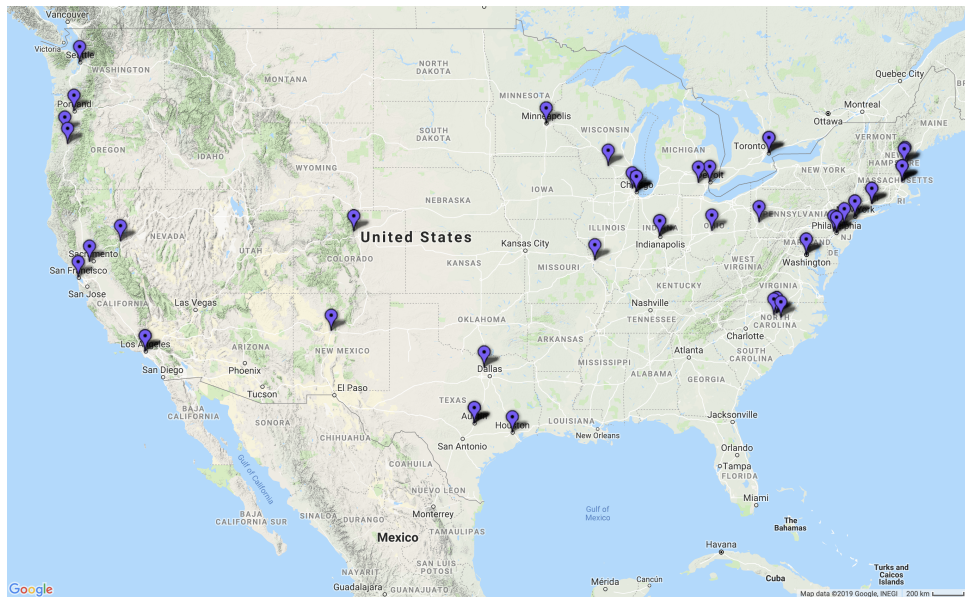


FIGURE 6.1: A map of DataRescue events between December 2016 - June 2017. The map was created by compiling data about DataRescue events from an archived version of the Penn Program in Environmental Humanities DataRescue web page (Penn Program in Environmental Humanities, n.d.).

to support common goals (Star and Griesemer, 2015, p.195). On the other hand, Gieryn (1983) uses *boundary work* to describe scientific practices that operate at a level of exclusion – where professionalisation works to foreclose and marginalise different forms of knowledge production, often to maintain autonomy and authority. Rather than attempt to reconcile this potential contradiction, I draw on both of these concepts to explore DataRescue as a system of boundary objects and highlight two ways that web archiving simultaneously *produces* and is *produced by* politics.

First, environmental information and data is itself a boundary object. In the case of DataRescue (and beyond), the preservation of access to this data became the goal that mobilised a diverse set of actors, skills and expertise to participate in political action in anticipation of an ‘anti-science’ administration. Here, the political climate and threat of the incoming administration to science data and environmental protections (described through the crisis discourse, above in Section 6.2) created, in the words of one informant, ‘a solidarity movement’ dedicated to intervening in the future of environmental data accessibility.

Second, in order to satisfy the diversity of needs and concerns brought by different actors, EDGI and the DataRescue project had to ‘translate’ these multiple perspectives to design sociotechnical interventions to achieve the collective goal of ‘rescuing data’. The DataRescue workflow and associated applications constructed by EDGI and partners can therefore also be seen as boundary objects that became negotiated (then locally standardised) representations of the priorities of different stakeholders that enabled the work of event-based web archiving. Consequently, despite the successes

of the events (discussed below in Section 6.4), DataRescue also highlights the politics of web archiving as revealed through the challenges of boundary work that attempts to both translate and lay claim to particular forms of expertise and practice.

The following Section 6.3.1 first explores the ways that environmental data mobilised multiple publics, before then considering the politics of the collaborative interventions (as toolkit, workflow and software applications) that enabled DataRescue in practice.

6.3.1 Mobilising Publics and Expertise

The ‘guerrilla archiving’ (GA) event in Toronto was the first of its kind: a public, large-scale, grassroots, community-based event that combined participation in the End of Term crawl and other efforts to download ‘difficult to crawl’ data. Despite a snow storm, the one-day event attracted around 150 attendees¹⁴ and a swarm of media attention, including coverage from the *Washington Post* (Dennis, 2016a), the Canadian Broadcasting Company (CBC) (Mortillaro, 2016) and *Vice* (Koebler, 2016), amongst others. Here I want to explore the ways that environmental information and data acts as a boundary object that mobilised participants with a range of expertise to both contribute to the DataRescue project and ‘lay claim’ to a set of archiving activities.

EDGI members have themselves reflected on how DataRescue and the formation of the EDGI organisation benefitted from the wider political context that surrounded the election of Trump – where, following the election there was a climate of political protest in the US and abroad in resistance to the incoming administration (Vera et al., 2018). During Trump’s first few months in office alone, several large-scale protests occurred across the US, including those championing women’s rights,¹⁵ refugee, migrant and immigration rights to travel,¹⁶ and in support of public science and evidence-based policy-making.¹⁷ Drawing on *political opportunity structure* theory (Kitschelt, 1986), Vera et al. (2018, p.517) describe the ways that EDGI’s pre-existing

¹⁴<https://enviroadatagov.org/datarescue> (visited on 21st May 2019)

¹⁵The international Women’s March occurred the day after Trump’s inauguration in January 2017 (including over 650 separate marches in the US alone), marking the largest protest in the history of the US (Chenoweth and Pressman, 2017).

¹⁶Also in January 2017, whilst occupying many international airports across the US, an estimated ‘thousands’ of people protested the passing of a (Presidential) Executive Order that put a freeze on refugees and banned travel from seven Muslim-majority countries (Gambino et al., 2017).

¹⁷The March for Science occurred on Earth Day in April 2017 across more than 600 cities across the US and the globe (Fleur, 2017), with goals of celebrating science for the public good and advocating for evidence-based policy making. <https://web.archive.org/web/20170318183426/https://www.marchforscience.com/mission-and-vision> (visited on 21st May 2019, archived on 18th Mar. 2017)

networks through distributed partner organisations like Public Lab¹⁸ and their ‘intellectual networks of STS scholars’, worked as organisational structures that also configured social mobilisation. These networks, combined with the urgency of the wider political climate facilitated a coalition of ‘diverse actors that do not often interact with each other’ (Vera et al., 2018, p.524) to coalesce around a social movement dedicated to enacting environmental data politics.

This sentiment is reinforced by several members who reflected on how DataRescue created a broad coalition (or in the words of Bernard, ‘a solidarity movement’) with a range of organisations and activists – including for example, Civic Tech Toronto, Data Refuge, Climate Mirror, Project Azimuth, the Union of Concerned Scientists and others – who each brought particular sets of methods, expertise, values and goals to the project of web archiving. Again, there was a feeling that the political climate brought together groups that ‘wouldn’t normally work together’, but also that it presented a ‘golden opportunity’ to seek creative interdisciplinary interventions:

“So for me [...] I was like this is the chance for us to build the library that doesn’t burn. This is a chance for us to establish consensus across professions and roles that normally wouldn’t want to work together. There is a golden opportunity here. And though this all feels really chaotic and we’re all feeling really upset, this is actually an incredible moment. This is a chance for us to do something really cool.” (Luke)

Here Luke makes reference to cautionary tale of the Library of Alexandria whilst championing the opportunities afforded this coming together of people and expertise to build consensus around community-held data. The intersection of concerns surrounding the preservation of environmental data extended to the many individual participants who attended the DataRescue events, some of which became members of EDGI and subsequently organised their own events. As Lamdan (2018, p.234) notes, DataRescue was ‘organic, innovative and optimistic, driven by the idealistic notions of saving and valuing government data’; a sentiment neatly represented in Augustine’s observations about the ‘weird way that people cared about data’:

*“[...] it was the public, immediate, weird way that people cared about data in a new way that felt special about the DataRescue events. I don’t think people expected **that** many people to show up in Toronto. In Canada. In December. And then the Philly one was big and Ann Arbour was big. And I was at Ann Arbour and people showed up and had bought a hard drive, because they were like ‘I don’t know, I’m not technical but is this what I’m supposed to do? This is where the data goes? But I’m going to help’, right? And maybe slightly in jest but also **so** earnestly. And that sentiment felt*

¹⁸The Public Laboratory for Open Technology and Science (Public Lab) is a community and 501(c)(3) non-profit corporation that works to ‘democratize science to address environmental issues that affect people’ through the development of low-cost and open source environmental monitoring. Public Lab serves as fiscal sponsor of EDGI. <https://publiclab.org/about> (visited on 21st May 2019)

novel, or important not to ignore which was people who were not research scientists or data managers, or web archivists or librarians or I don't know people [...] who should know this shit—caring about this preservation, that seems pretty special. But I mean it was tied to this political climate and level of uncertainty and fear.” (Augustine)

Augustine first reflects on the level of participation and directly links it to the political climate and uncertainty of the post-election atmosphere. Recalling the Ann Arbor event, Augustine points to the uniqueness of the atmosphere, and places particular emphasis on participation from volunteers who had no previous connections to either digital preservation or specific research datasets. Here environmental data acts as a boundary object that extends beyond the scholarly and professional disciplines already working in digital preservation to include a wider set of publics concerned about environmental protections and the role(s) of science, the state and activism in ensuring those protections. DataRescue became a means for the creation of a ‘novel’ form of civic engagement; a form of activism that EDGI scholars (Currie, Donovan and Paris, 2018; Currie and Paris, 2018) have argued draws on broad traditions of both *archival activism* (Flinn and Alexander, 2015) and *data activism* (Milan, 2016).¹⁹ Currie, Donovan and Paris (2018) make the case for DataRescue as an example of *activist data archiving*, providing insights into the ways that grassroots data archiving projects become organising sites that mobilise decentralised means to bridge ‘experts and lay publics’ in the service of social justice aims. They expand on this work to position data archiving as a form of political activism to highlight ‘the power of archival work to safeguard politically vulnerable information’ (Currie and Paris, 2018).

Other EDGI members have further reflected on the ways that DataRescue enabled a broad public to participate in political action. Here, web archiving enables a form of activism that EDGI member Gary described as a way of doing something ‘for the greater good’ that was ‘less confrontational’ than attending a protest. Another EDGI member, Julian, talked about being initially motivated to participate by both a concern for the ‘disappearance’ of scientific data and a prior awareness about the precedent set in Canada during the Harper years. Julian spoke about the ways that the DataRescue event itself expanded their concerns about data availability, as they learned more about the potentially vulnerable environmental and physical sciences datasets that were nominated by scientists for archiving. Below, Julian and Bernard both express how their involvement enabled them to channel their concerns and apply their own expertise to ‘do something valuable’ in an effort to effect political change:

¹⁹Data activism focuses on the ‘reappropriation’ of statistics and data power (Bruno, Didier and Vitale, 2014) for the purpose of recognising the ‘political dimensions of data’ through the generation of new data or statistical analyses that challenge official representations (Currie, Donovan and Paris, 2018, p.67-68). Archival activism (and similarly, activist archiving) encompasses both the activities of archivists who ‘seek to creatively document political and social movement activism as well as those projects which engage with archives and the archival process as part or in support of political, human right and social movement activism’ (Flinn and Alexander, 2015, p.329).

“[...] although I generally abhor hackathons, I kind of felt like it seemed—from what I understood of [DataRescue]—they were scoped in a way where it might be useful and I could go and do something valuable about a problem that I was super concerned about. And that I could also do something political that felt like it might be useful with my technology skills.” (Julian)

“[...] it felt like I had something to offer. And I’ll say I care about environment and climate and I’ve written about [it] a bit, but it’s not like that was obviously the thing that I cared most about. It was that this was a project—and this is something else to be said, it’s not to say that we’re the best people to do this, it’s that we’re the people who are doing it. And it felt like something that ought to be done and that I had something to offer too, and so it really filled a need for me because not having anything to do was very distressing. So I think that there were a lot of people in that situation that really just wanted to have something to do, some way to take action as a citizen, you know? To intervene in the transformation of the United States.” (Bernard)

Julian and Bernard position the goal of web archiving as a mobilising force for their own participation; a goal that they both note is political in its aims. Along with other EDGI members, they both reflect on the ways that web archiving enabled them to ‘take action’ and ‘do something political’ with their skills. Beyond the observation that Julian and Bernard believed they had valuable skills to offer the cause of DataRescue, Bernard’s comments also work to legitimise their involvement in web archiving, despite a view that EDGI were possibly not ‘the best people to do this’. This positioning and centring of expertise, can be seen as a form of boundary work that is the result of a broad public caring about and claiming a set of activities they may have otherwise not engaged in – a phenomena that Augustine attributes to the ‘political climate and level of uncertainty and fear’.

During DataRescue, web archiving becomes the place where the politics of expertise and practice are negotiated by participants from different fields to achieve the end goal of preserving access to government data. The significance of these politics extend to the role of the media in communicating the goals of DataRescue and mobilising participants to the cause. Describing how they first heard about the Toronto event, one EDGI member, Ethan recalled receiving a link to a *Washington Post* (WP) article (Figure 6.2) from a friend on Facebook that reported on various efforts taking place to preserve climate data (including the GA event, and other projects like Climate Mirror). Following Trump’s election, Ethan – who described themselves as ‘not very politically engaged’ prior to the election – began meeting with university friends and colleagues who had broad concerns about ‘institutional memory loss’ at federal agencies and the potential impact of both departmental re-shuffles and career staff departures due to the incoming administration’s ‘anti-science’ agenda. The WP article – in addition to accentuating the state of crisis – centred the role of ‘scientists’ in their

‘frantic’ and ‘feverish attempt to copy reams of government data onto independent servers’ (Dennis, 2016a). In an interview, Ethan and I discussed the potential ways that the media ‘messaging’ surrounding the event played a part in mobilising them to participate:

Ethan: *“[...] the messaging was right because it was scientists, so it kind of caught [my] eye at that time probably, even though I think that part really isn’t—I don’t know, I think it was important for the public perception, that it’s scientists but I don’t actually think that that was really important in terms of who was doing the work. The way it was done, maybe the way we wish we did it.”*

Jessica: *“Do you think that—did that draw you in though? The idea that it was being organised by scientists or being pitched to scientists?”*

Ethan: *“Probably to some degree. I don’t know, yea if they had just said librarians are doing this maybe [I] would have thought about it a little differently.”*

Although Ethan is reluctant to fully attribute their own participation to the messaging in the article, it is nonetheless clear that they believed it had an impact on the public perception of who was doing the ‘guerrilla archiving’. They went on to speculate about the effectiveness of the message, and whether or not their friend had sent them the article because of their ‘scientist connection’. This works to briefly highlight the significance of the discursive practices that surrounded the wider interpretations of ‘guerrilla archiving’, DataRescue and the ways that *who* was seen to be doing these activities had an impact on how support was mobilised. These discursive politics are further observed through Bernard’s reflections about the significance of describing the first event as ‘guerrilla archiving’:

*“[...] in some ways ‘guerrilla archiving’ has a different feel right? So there are two things about it. First it has archiving in the name so it explicitly lays claim to a certain kind of activity despite the fact that there were no archivists involved, right? And then it has this guerrilla word which is a little bit aggressive in some ways but certainly—so in some contexts it’s perceived as aggressive—it’s certainly oppositional. So it had that flavour of we are going to archive despite the fact that they don’t want us to. And that’s how we understood what we were doing. Although I think we didn’t really know who **they** were who didn’t want us archiving.”* (Bernard)

Bernard acknowledges the explicit politics embedded within the discursive practices surrounding the naming of the first event. The GA event worked to assert and legitimise a claim to a set of memory practices typically limited to the boundaries of professional librarians and archivists. The name also communicates a sense of covertneess and ‘opposition’; implicitly positioning the not-yet-in-office administration as

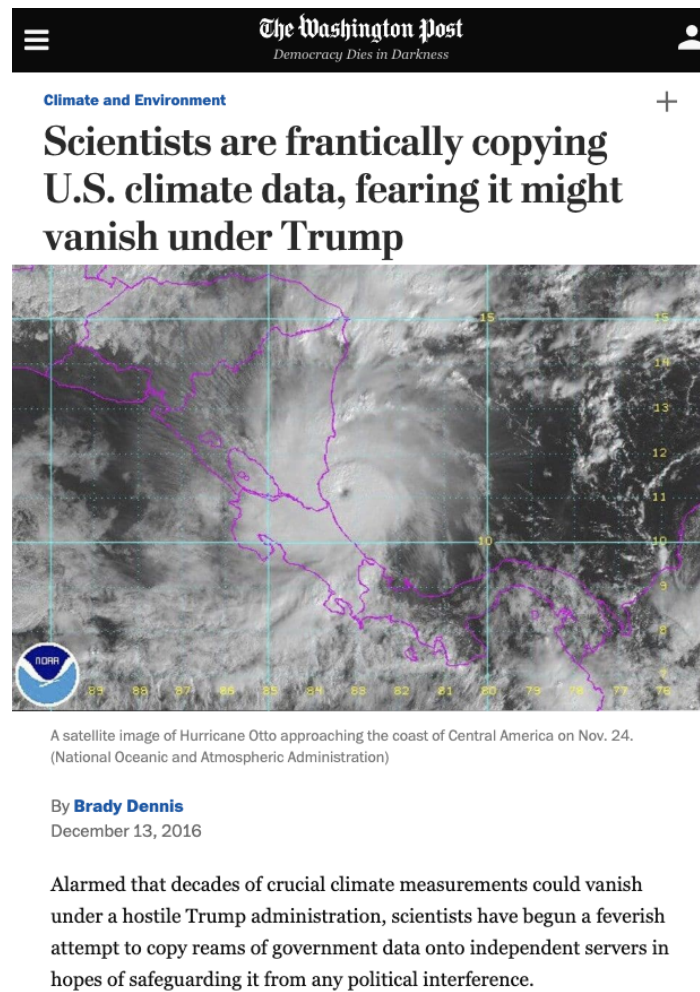


FIGURE 6.2: A *Washington Post* article that included coverage of the 'guerrilla archiving' event and other efforts to save environmental data in December 2016 (Dennis, 2016a).

opponents in the preservation of environmental information access. Although EDGI and the TRU brought specific sets of expertise to the organisation of the GA event – including, for example, skills in activist/community organising, data science, STS, environmental justice, history and information science scholarship – Bernard and others have noted the significance of EDGI's positioning outside of the professional expertise of librarians, archivists, or indeed web archiving circles of practice. This framing is reminiscent of other scholarship on the power of community archiving, where EDGI members and DataRescue participants worked to bridge different 'epistemic cultures' (Knorr Cetina, 1999) and boundaries of expertise to '[seize] the archive as an apparatus to legitimize the new forms of knowledge and cultural production in an economically and politically precarious present' (Eichhorn, 2013, p.4). Here the archive becomes a vehicle for political resistance – where resistance becomes a 'diagnostic of power' (Abu-Lughod, 1990) and sheds light on the power relations embedded in both the vulnerability of public data infrastructures and the professionalisation of memory practices associated with ensuring their stability.

6.3.2 Collaborative Interventions

In the lead up to the Toronto GA event, EDGI began collaborating with Data Refuge, a project based out of the Penn Program in Environmental Humanities (and the University of Pennsylvania Libraries). Data Refuge also formed in November 2016 ‘to draw attention to how climate denial endangers federal environmental data’.²⁰ Working with the model initially developed as part of the Toronto event, the groups collaborated on the further development of an open workflow for other institutions and community groups to host their own events. From January 2017, EDGI and Data Refuge members liaised with local DataRescue organisers to onboard and share information; communicating regularly through online community conference calls, email, Slack and GitHub. Events were organised by university libraries, environmental programmes and labs, civic technology groups, and others, with some organising multiple meet-ups to both participate in DataRescue and the technical development surrounding the harvesting of web pages and data. In addition to managing the emergence of EDGI as an organisation, during the peak of DataRescue activities (December 2016 - March 2017) EDGI members involved in DataRescue reported to have spent anywhere between *5-50 hours a week* organising events, coordinating the toolkit and software contributions, attending events across the US, developing protocols, creating documentation and generally managing the collaboration with Data Refuge.²¹

Between December 2016 and June 2017, the event model evolved over time, as work with local organisers combined with many hours of coordination, collaboration and research by EDGI members and others worked to create documentation and tools for community-based web archiving. Here, Bernard describes the initial model used at the GA event, which reflects some of the early motivations behind grouping archiving tasks into separate areas based on both the skillsets of participants and a perceived difference between archiving web pages and datasets:

“We started to develop a taxonomy in preparing for that event when we were trying to think about who are the different classes of people who are going to participate. We had these different areas. There was a strategising or political organising area, there was a seed generation area which is where most of the people were, and this thing that we called ‘hackers corner’ where there were people trying to figure out how to extract datasets from web interfaces. So we called that the ‘hard to crawl’ or ‘uncrawlables’ section. And so we were totally naively working our way through a set of problems that of

²⁰<http://web.archive.org/web/20190304214339/https://ppeh.sas.upenn.edu/experiments/data-refuge> (visited on 22nd May 2019, archived on 4th Mar. 2019)

²¹These figures are drawn both from my own interviews and an internal 2017 ‘EDGI Lessons Learned’ collaborative presentation drafted by members (across the different working groups) in an effort to document the first eight months of the EDGI organisation.

course web archivists have thought about. But since there were no web archivists there, you know I mean we just had to blindly work our way through these issues.” (Bernard)

The basic structure of the workflow evolved to eventually include five optional ‘paths’ or tracks that contained different types of activities within them, including tracks that marshalled different forms of expertise to systematically map the government webspace, archive websites and manage the processes associated with ‘difficult to crawl’ datasets. Drawing on the workflow documentation designed by EDGI, Data Refuge and organisers, the following briefly outlines each track:²²

Path I: Surveying – creating ‘agency’ and ‘sub-agency primers’ for mapping the web presence of government agencies, offices and programmes, including researching their background, incoming administration policies and appointments, and potential risks of deletion associated with particular sites and data

Path II: Website Archiving – using agency, sub-agency primers and a custom Google Chrome extension to contribute seeds to the End of Term project, as well as work to identify difficult to crawl datasets

Path III: Archiving More Complex Datasets – researching datasets discovered in Path II and writing scripts to scrape and download them. Coordinated by the *Archivers.space* event workflow management tool, including the following roles:

- **Researching** if and how particular datasets could be harvested
- **Harvesting** difficult to crawl datasets using custom scraping scripts
- **Checking and Bagging** to inspect and confirm the integrity of datasets, then ‘bag’ the data using the *BagIt* file format to generate checksums and deposit them in the Data Refuge CKAN repository²³
- **Describing** datasets in the CKAN repository, increasing potential usability through basic metadata and making data publicly accessible online

Path IV: Storytelling – using social media and storytelling kits developed by Data Refuge to collect stories about the DataRescue event and participants, including a focus on the importance of climate and environmental data in everyday life

Path V: Looking Beyond DataRescue Events (*the Long T(r)ail*)²⁴ – a space dedicated to roundtable discussions for the future of various EDGI and Data Refuge projects, including:

- *EDGI’s Next Steps in Tech Development* to discuss community building and strategise the potential for distributed, community-based public data stewardship
- *Data Refuge Built into a Libraries Network* to build a network of research libraries in the US to systematically ‘pull’ government data
- *Data Refuge’s Longer Path: Three Stories in Our Town* project to work with communities to document how they use environmental and climate data

²²This model and the text included here summarises and draws on multiple online sources in EDGI’s archived GitHub repositories: <https://github.com/edgi-govdata-archiving/DataRescueTEMPLATE> (visited on 14th May 2019) and the DataRescue Workflow: <https://datarefuge.github.io/workflow> (visited on 14th May 2019)

During the peak of DataRescue, the workflow itself can be seen as a central *system of boundary objects* that became a continuously negotiated and temporarily stable ‘bridge’ (Star and Griesemer, 2015, p.194) between EDGI, Data Refuge, local event organisers and the End of Term project partners. Event organisers locally adapted the workflow to suit their own needs whilst using the tools to contribute to the wider project activities. The tools and workflow enabled EDGI to take advantage of and extend the End of Term project infrastructure – through the Chrome extension, agency primers and extensive volunteer labour – whilst contributing to the seed nominations. Each of the paths reflects the priorities of particular collaborators that brought different sets of concerns and professional expertise to the problem of facilitating the processes behind archiving websites and data. They required translation and buy-in from a diverse set of stakeholders; where EDGI, in particular was driven by a commitment to horizontal governance and decision-making. DataRescue coding work during this time was often facilitated by what Luke referred to as a mode of working that championed ‘rough consensus and working code’ to enable the technical development to happen in real-time with the event coordination:

“I was like OK we do have consensus around that we need backups and we need backups as soon as possible. And we don’t know how to do that but we’re going to try. And the most exciting thing about that phase and the time that I think a lot of people refer to as ‘the DataRescue time’ is in the beginning there was this real—OK, any given solution will be accepted as long as it fits. [##] has this incredible phrase ‘rough consensus and working code’. If there’s rough consensus around a topic and the little script you wrote does the thing, we use it. And that tended to be the ways that things moved forward at the time.” (Luke)

Below, I briefly explore the ways that both website and data archiving were negotiated through two components of Paths I/II and III of the DataRescue workflow: seeding the End of Term Crawl and archiving ‘uncrawlable’ data. These two interventions help to highlight the ways that DataRescue both broadened the communities of expertise involved in archiving websites and data in practice, as well as challenged organisers to translate their expertise and goals into working software and protocols that shaped the who, what and how of web archiving was accomplished.

²³CKAN (the Comprehensive Knowledge Archive Network) is open source software for managing and publishing open data online: <https://ckan.org> (visited 21st May 2019). Data Refuge setup a CKAN portal to house data harvested during DataRescue here: <https://www.datarefuge.org/dataset> (visited on 18th May 2019)

²⁴In different DataRescue documentation and discussions with EDGI, this path was interchangeably referred to as both the ‘Long Tail’ and the ‘Long Trail’.

Seeding the End of Term Crawl

In addition to the workflow documentation itself, the EDGI toolkit included guidance to encourage local organisers and attendees, through the promotion of a code of conduct, to create events ‘aimed at fostering an inclusive and enabling work environment’ (Dillon et al., 2019, p.7).²⁵ EDGI produced documentation on web crawling, including guides on the Internet Archive web crawlers (Environmental Data & Governance Initiative, 2016a,b) used for the EoT project and training videos that walked organisers through the basics of using the software and primers as part of the DataRescue workflow. These were designed by and for volunteers who had little-to-no previous experience with web crawling, however provided a mechanism for DataRescue participants to utilise and extend the EoT infrastructure for archiving federal websites. This work drew on the goodwill of volunteers to research and manually nominate seeds, as well as the wider expertise of researchers who had knowledge pertaining to public science data and information, environmental policy and the breadth of resources potentially vulnerable to removals.

As discussed in Section 6.2.2, the Toronto GA organisers identified the EoT project early on as a key mechanism for channeling participation at the event. However, as Augustine explains, the existing End of Term nomination tool (developed and hosted by the University of North Texas) did not scale to the size and use-case of the events model, as it was not designed for repetitive use:

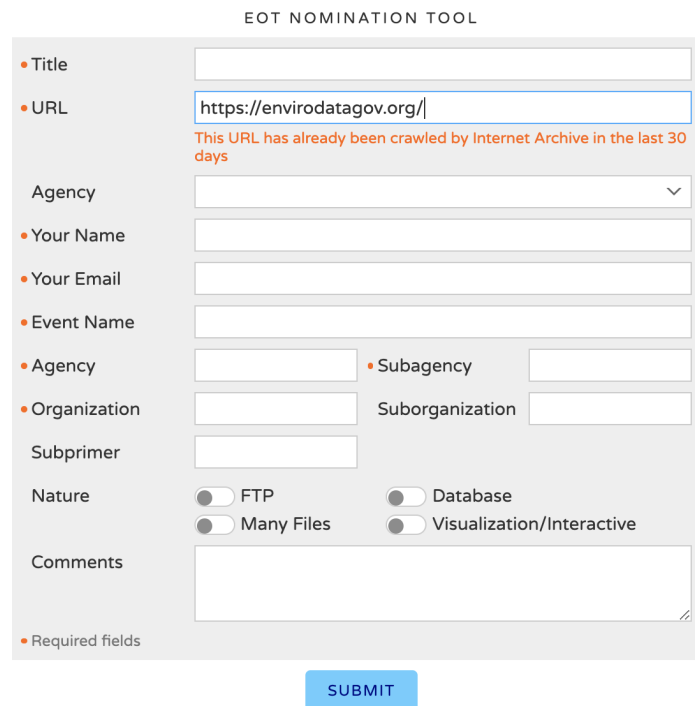
“[the UNT bookmarklet] was great but it worked for one thing. And then you had to re-enter all of your information the next time. And we were like, OK—because immediately, like three in, you would be so annoyed—I was annoyed. So we were already back in December [...] thinking about—we already knew that nominating to the Internet Archive was a great fit for everyone’s technical background but then how to make that process easier.”
(Augustine)

In order to contribute to the EoT project, GA organisers solicited help from Civic Tech Toronto, a local civic technology group who run a weekly open ‘Hacknight’ for Toronto community members to pitch collaborative projects with volunteer developers.²⁶ Two Civic Tech Toronto volunteers collaborated with the GA organisers to code the Chrome extension for the first event.²⁷ The extension operated in the Chrome browser and enabled users to input basic information associated with the title and URL of the targeted site, as well as details about the submission, including details pertaining to the user and the DataRescue event (see Figure 6.3 for a screenshot of the tool in its current state). Once fully developed, the Chrome extension

²⁵At the time of writing, the toolkit documents developed by EDGI (including the code of conduct) are still hosted on their website at: <https://envirodatagov.org/datarescue> (visited on 14th May 2019)

²⁶<http://civictech.ca> (visited on 15th May 2019)

²⁷The code for the Chrome extension: <https://github.com/edgi-govdata-archiving/eot-nomination-tool> (visited on 18th May 2019)



The screenshot shows a web form titled "EOT NOMINATION TOOL". It contains several input fields and a "SUBMIT" button. The fields are: Title, URL (with a warning message: "This URL has already been crawled by Internet Archive in the last 30 days"), Agency (dropdown), Your Name, Your Email, Event Name, Agency (text), Subagency (text), Organization (text), Suborganization (text), Subprimer (text), Nature (with radio buttons for FTP, Database, Many Files, and Visualization/Interactive), and Comments. A legend at the bottom left indicates that fields with a red dot are required. The "SUBMIT" button is blue and located at the bottom center.

FIGURE 6.3: A working screenshot of the EDGI Google Chrome extension used to submit seeds to the 2016 End of Term project, as part of the DataRescue events. If the URL had recently been archived, the user received a message to avoid duplication of effort.

enabled users to also enter government sub/agency codes (discussed below) used by EDGI to systematically keep track of captured websites and datasets that they identified as vulnerable. In order to avoid duplication of efforts, the extension also queried the Internet Archive's CDX API to confirm the URL had not been archived within a specified time window. Once entered, the extension submitted the seed information to the UNT bookmarklet for focused crawling, as well as to a Google form and spreadsheet centrally managed by EDGI.

Following Toronto, the EoT seeding activities at DataRescue were organised via the creation of *agency primers* and *sub-agency primers* that aimed to systematically map federal agencies, sub-agencies (offices), programmes and projects to their relative websites and pages. These primers eventually served several purposes for EDGI. They were used to coordinate and systematise the selection of nominations and to lower the technical barrier to entry for seeders using the Chrome extension. The extension and primers also became a tool for 'peer learning' that enabled volunteers to both research and become informed about the organisational structures of US federal environmental agencies and programmes. The primers also later became another mechanism to prioritise the selection of sites saved via *Versionista*, the tracking software used in the early stages of EDGI's Web Monitoring project. Speaking about how the primer system emerged after Toronto, Ethan reflects on their desire to make the process of nomination more systematic:

“[...others were] saying hey this was good but we need to find a way to organise on a high level what we start tackling. So that’s how we made these agency primers. I remember exactly where I was in [##], in some cafeteria being like damn it now we have to do this. Because I didn’t see another way and we were just sitting there talking about how we make this systematic and the only way we could think about it was actually using the government’s own structures to organise. So we used org charts and made an agency office database [...]. It’s basically just a relational database that attributes a code to a particular office. And it’s a hierarchical structure so that there’s an office, a higher office, an agency and a department. And that way you could relate a website to a particular part of the government. Even though websites are not always that simple. But we were like well if we span all the agencies and all the offices then we kind of span the government, at least in some measure of it.” (Ethan)

Ethan describes using the the organisational structure of government agencies to map their relative web presence. Hosted on Google Docs, the primers used basic templates (see Figure 6.4 for an example of an agency primer) to outline the offices contained within each targeted government agency, including the background of the department and a risk assessment regarding potentially vulnerable programmes and datasets within each sub/agency. Vulnerability for website/data change and loss was assessed based on key appointees, nominees and statements by the incoming administration.²⁸ The primers created a map for ‘seeders’ to follow at DataRescue events, guiding users to explore each agency and office website, recording URLs along the way. Different DataRescue organisers and volunteers would ‘claim’ certain offices and programmes (and associated sub-primers) during their events in an attempt to reduce duplication of efforts across DataRescue seeding efforts.

The Chrome extension and agency primers built to systematise the seeding efforts broadened the community of participants to include both volunteers with domain-specific knowledge about environmental science data and those without the technical skills to archive websites and data at scale. The tool enabled the GA event (and the DataRescue events that followed) the benefits of contributing to the EoT project in a distributed way and established a mutually beneficial relationship with existing (and long-running) web archiving efforts and organisations like the Internet Archive and the University of North Texas Libraries. Beyond assisting with the initial development of the Chrome extension, Bernard reflected that Civic Tech Toronto became instructive for EDGI (especially during the early stages of the organisation) on ways to to model community work around feminist and anti-racist principles of inclusivity to enable open source technology for public good. Through this collaboration,

²⁸EDGI provided guidance in a Google Doc entitled ‘How to Create Primers’ that was linked to the DataRescue Workflow that gave examples and guided primer ‘surveyors’ through the creation of the document.

Agency Office Code: [1-0-0-0]

Environmental Protection Agency (EPA)

Version 1.3

Main Agency Primer

I. Organization Sub-Primers

This is a growing list of primers on particular offices or organizations within the Environmental Protection Agency. Each sub-primer contains specific links for use in archive-a-thons, as well as guidelines to relevant data sets.

- [Office of Enforcement and Compliance Assurance \(OECA\)](#)
- [Office of International and Tribal Affairs \(OITA\)](#)
- [Office of Environmental Information \(OEI\)](#)
- [Office of Land and Emergency Management \(OLEM\)](#)
- [Office of the General Counsel \(OGC\)](#)
- [Region 6/ Dallas](#)
- [Office of Chemical Safety and Pollution Prevention \(QCSPP\)](#)
- [Office of Air and Radiation \(OAR\)](#)
- [Office of Water \(OW\)](#)

II. Background

The mission of the [Environmental Protection Agency](#) is to protect human health and the environment through the development and enforcement of regulations. The EPA is responsible for administering a [number of laws](#) that span various sectors, such as agriculture, transportation, utilities, construction, and oil and gas. In the [budget for FY 2017](#), the agency lays out goals to better support communities and address climate change following the President's Climate Action Plan. Additionally, the agency aims to improve community water infrastructure, chemical plant safety, and collaborative partnerships among federal, state, and tribal levels.

III. Transition Team for EPA:

(Note that all members are volunteers except Myron Ebell and David Kreutzner, who are "privately funded".)

Prior to the inauguration, the EPA was managed by the administration's transition team. With Myron Ebell at its head, the transition team suggested a shift away from the climate programs, specifically the Clean Power Plan and President's Climate Action Plan, while refocusing on [air and water programs](#). The transition team and new administration plan to focus on ensuring [cheap, abundant energy](#), regardless of its carbon intensity.

The EPA transition team was made up of two people from the Competitive Enterprise Institute, a libertarian think tank dedicated to free enterprise and limited government. There was one person from each the libertarian think tanks the Independence Institute and Caesar Rodney Institute, and one person from each the conservative think tanks The Heritage Foundation and The Federalist Society. There was one person from the coal funded Energy and Environment Legal Institute and two previous congressional staffers. The appointed head of the EPA is

FIGURE 6.4: A screenshot of an EPA Agency Primer (Version 1.3, page 1 of 2) used as part of the DataRescue workflow. Includes links to the agency's sub-agency primer documents, the background of the EPA, further information pertaining to relevant appointments to the 2016-2017 Trump transition team and a risk assessment for particular agency programmes.

EDGI members also identified and established a working relationship with the Internet Archive, an organisation seen to have complementary organisational values and with whom they still work in the context of the Web Monitoring project.

During and after the Toronto event, EDGI volunteers invested heavily in developing the DataRescue workflow with Data Refuge, which emphasised both non-technical and technical contributions to the project – expanding on the different tracks to enable participation from people with varied skills and expertise. Nonetheless, although the Chrome extension was pitched to people with little-to-no technical skills, in practice it revealed complexities of communicating web archiving (and DataRescue goals) to both non/technical participants (at the event and within organising team at Toronto).

Archiving the ‘Uncrawlable’

The conceptual and technical differences between preserving web pages and datasets that emerged from the Toronto event was said to have acted as a ‘social organiser’ for both the DataRescue workflow and EDGI’s archiving work, more generally. In Toronto, ‘hackers corner’ became a space largely occupied by technologists and people with coding experience (data scientists, civic technology experts, and more) who contributed to scouting and developing ways to map and download environmental data from the federal webspace. Implicit in this separation is the different ways that ‘data’ was framed by the workflow and its creators. Here, data is conceptualised based on the tools and technologies needed to extract it from the Web. In the case of basic HTML, the End of Term crawlers (using Heritrix) were sufficient for automating the collection. Other types of data, like those illustrated in Allen et al.’s (2017) graphic (Figure 6.5), present known problems for crawlers like Heritrix, and required further investigation to design collection mechanisms. Data stored in databases with queryable web front-ends, embedded content and other large-scale directories of packaged datasets became the target of custom scripts and efforts to semi-automate and extract data from web interfaces. To organise these efforts, initially DataRescuers used centrally managed Google spreadsheets to keep tally of URLs and datasets that required researching custom harvesting techniques. URLs and datasets were given a unique ID (UID) and researchers and harvesters would ‘claim’ the ones they were working on during the event by placing their name in the corresponding spreadsheet column. This portion of the workflow development process benefited from what Luke described as ‘a lot of iteration’:

“So the tech development at the time in Philly—well it was developed largely at—well kicked off at the Toronto event where everybody just chucked up a Google sheet and a lot of the iteration, and this was incredible—software start-ups could learn a lot from this phase—where it was just chuck it on a Google sheet, do it all with existing tools and iterate by changing the names of the columns. So what does this thing mean, just put your GitHub username in this cell if you’re working on this thing. And that was like our checkin/checkout system. And iterating on that rapidly during the day. And particularly while the Philly event was happening we were constantly being like—like OK, we should all use this thing, and [we] very quickly wrote a script for like UID generation and copy-pasted a column of 600 new UIDs, and was like OK this will do that.” (Luke)

Luke describes one of the ways that DataRescue partners were collectively designing the workflows for harvesting ‘difficult to crawl’ datasets through the use of existing tools that enabled community members to rapidly modify aspects of the workflow. However, as the number of events, participants and targeted resources increased, capacity issues with this portion of the workflow were encountered. Problems emerged

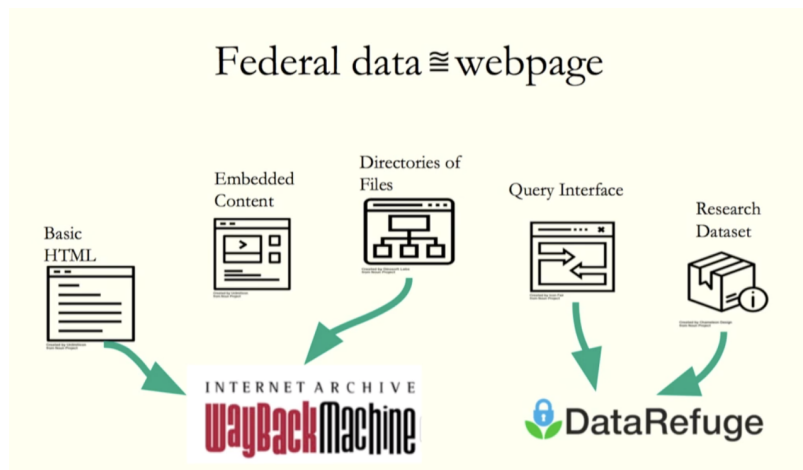


FIGURE 6.5: Spectrum of data archiving at DataRescue, including how efforts directed data capture towards different repositories (Allen et al., 2017).

at the Ann Arbor DataRescue which drew around 300 people (over the course of two days), where the online spreadsheets ‘stalled’ under the number of simultaneous connections, resulting in data loss and a duplication of effort (when volunteers didn’t realise others were working on the same tasks).

To mitigate the issues associated with the scale of participation in the events, over the course of ‘a series of all nighters’ an EDGI member initially developed *Archivers.space*,²⁹ a web-based application to manage the workflow for datasets and scale practices that were hitherto being ‘iteratively’ managed via spreadsheets. The home page of the app enabled organisers to visualise the work of each DataRescue event (Figure 6.6). The app also introduced ‘a mutual exclusion lock’ that allocated and tagged URLs with universally unique identifiers (UUIDs) and prevented them from being checked out whilst others were working on them. Built to ‘exactly match the community-designed workflow’, the app became an event management tool that walked volunteers through the steps of researching and harvesting (Figure 6.7), as well as tracking datasets as they were elsewhere ‘bagged’, described and deposited in the Data Refuge CKAN repository. As a mechanism for manually documenting the data’s origins, the app helped users add metadata about the datasets and harvesting techniques used, and created template (zipped) file directories for volunteers to compress the data alongside the custom harvesting scripts used. However, whilst the app addressed the practical problems of scaling the overall coordination of researchers and harvesters, parallel tensions emerged surrounding epistemological differences over the nature of provenance and the community’s capacity to flexibly iterate on the workflow.

From the early stages of DataRescue organisers were concerned about web archiving issues related to data reliability and provenance that both informed the development

²⁹<https://www.archivers.space> (visited on 31st May 2019)

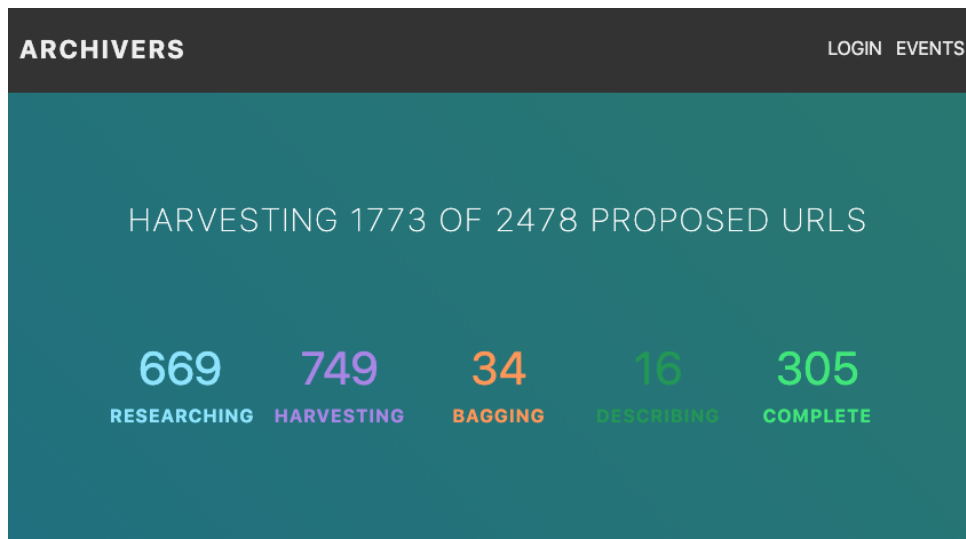


FIGURE 6.6: Screenshot of the home page of the Archivers.space application. The application showed the total number of items in each phase of the DataRescue workflow.

of practices and produced particular tensions around paths forward. The activities contained within (what would eventually be called) Path III reflect the priorities of digital preservation best practices particularly concerning issues for ensuring the provenance of data, data ‘fixity’ and establishing a trusted ‘chain of custody’. Data Refuge collaborators with backgrounds in maintaining research data infrastructures have elsewhere reflected on their central concerns over ensuring the lifecycle and future use of harvested data by researchers. Here Allen (2017) emphasises and reflects on embedding ‘chain of custody’ as a core value of the DataRescue workflow:

“[...] By documenting [the chain of custody] – where the data comes from originally, who copied them and how, and then who and how they are re-distributed – the Data Refuge project relies on multiple checks by trained librarians and archivists providing quality assurance along every link in the chain. Consider this extreme case: What happens if an original dataset disappears, and the only copy has passed through unverified hands and processes? Even a system that relies on multiple unverified copies can be gamed if many copies of bad data proliferate” (Allen, 2017).

Allen makes the case for ensuring the provenance of harvested data, describing concerns that arise when the verifiability of scientific data is called into question. This is increasingly central to the discussion of climate data, in particular, where data have been subject to various forms of scrutiny from a spectrum of stakeholders that range from efforts to participate in scientific knowledge production to those that are

ARCHIVERSLOGIN EVENTS

<http://epa.gov/airquality/greenbook/index.html>

Priority10

71F3E36D-EF73-4E10-9019-010635FE005A

status: harvest

URL Status: Checked out by at Fri May 19 2017 15:52:05 GMT+0100 (British Summer Time)

Info

Agency:

Event:

Title: Nonattainment Areas for Criteria Pollutants

Crawled by Internet Archive: Yes

Description:

Research

☒ Check here when this stage is complete

Page Title:

Purpose / Significance of Data

Optional. Ideally provided by a domain specialist or user of this kind of data.

☐ Do not harvest. All data is small, unstructured, and on a page crawlable by the Internet Archive.

☐ Page contains dynamic content (e.g., links loaded by JavaScript).

☐ Page contains interactive visualizations.

☒ Data is accessible in structured file(s) that can be directly downloaded.

☐ Data is accessible over FTP.

☐ Data is accessible using a documented public API.

☐ Data is only accessible using search queries in a web form.

Recommended Approach for Harvesting Data

If this will be handed off to someone else to harvest, pass on any useful info here.

File Formats

such as: PDF, CSV, XLSX, JSON, HDF5

FIGURE 6.7: Screenshot of an item in the Archivers.space application. The application allowed DataRescue organisers and participants to track each step of the workflow. This particular item has gone through the 'research' phase and is waiting to be harvested.

more representative of political and ideological strategies for spreading misinformation (Edwards, 2013, pp.425-426).³⁰ With this in mind, the enactment of strategies to avoid future controversy over the origins and validity of scientific data – particularly in the context of grassroots, non-institutional forms of web archiving – reveals inherent tensions at the boundaries of expertise. In response to a question about the role of provenance in the DataRescue workflow for archiving data, Luke reflected both on the problem of provenance and the distinctly different ways that participants framed the solutions:

“Yea I mean it’s just too many people touching the data. How do you establish a chain from—these are bytes on a website at a point in time and how do I draw a consistent and reliable through line from that to community-held data? And how can you prove that every step of the way that data hasn’t been tampered with? And that’s the standard. If you’re going to turn around and give that output to climate scientists, that’s the standard. So we talked about that a lot and the librarians were on to this from the get-go.

The thing that drove the technicians nuts is that the librarians believed that the provenance arose from a human procedure whereas all of us on the technical side were like, no no no this arises from keeping the humans out of this procedure. And bringing them into a system where they are administering data as it’s moving, not touching data directly. And that’s in direct tension with a grass-roots movement where it’s like the whole point is that they should touch data directly.” (Luke)

Here, onto-epistemological differences between how provenance is both conceptualised and operationalised in the context of DataRescue represents an intersection or boundary between different social worlds and forms of expertise that have bearing on the ways that web archiving is done. The above reveals two intertwined points of contention about the who and how of web archiving. The first point is around how to ensure verifiable provenance in practice. Above, Luke disputes the characterisation that provenance should be enabled through ‘a human procedure’, advocating for technical mechanisms for ‘keeping humans out’ of the process. In conversation, Luke went on to frame the problem as one that should be seen through the computational lens of ‘repeatability’, rather than a need to document ‘whose hands have touched it’, emphasising that in the context of DataRescue they actually needed tools that didn’t yet exist.

The second point is around who should perform these mechanisms of ensuring provenance. Whereas Allen (2017) emphasises the role and inclusion of experts (‘trained librarians and archivists’) in the process of facilitating the capture of data provenance,

³⁰ A prominent example of the latter is the so-called ‘hockey stick controversy’ which embroiled climate scientist Michael Mann and colleagues in years of public dispute and critique at the behest of climate change denier Stephen McIntyre (Edwards, 2013, pp.422-426). Here, the rhetorics of ‘open data’ and ‘open science’ was weaponised with powerful effect, producing and fuelling an on-going public debate around scientific methods, truth-making and the spread of misinformation for political gain.

Luke identifies an inherent problem between the ways that ‘keeping humans out’ of the process is in direct tension with a grassroots approach to web archiving. The implication is therefore that by removing activists from the processes of seeking out, downloading, describing and depositing data, a distance is created between people and data that is potentially in contradiction with the goals of activist web archiving. Whilst reiterating the performative view that web archiving is intrinsically tied to action, this observation also raises questions tied to the operation of power and legitimacy claims surrounding who (and in what context) web archiving should be performed. Here, web archiving practice is essentially linked to and framed through the lens of particular professional objectives embedded within the act of web archiving, in this case to ensure provenance.

Further issues surrounding expertise and web archiving extend to the ways that EDGI proceeded to model technical paths forward for capturing web resources not included in the EoT crawls. From December 2016, with a shortage of web archiving technical expertise, an impending sense of urgency, and the scale and pace with which the DataRescue movement escalated, some members described the impacts on their abilities to interpret and make use of existing web archiving tools in use elsewhere. EDGI members investigated other tools, including Heritrix, the crawling software used by the Internet Archive and by other International Internet Preservation Coalition (IIPC) members in national libraries. However, speaking about Heritrix, one EDGI member, Luke reflected that during the course of DataRescue they not only felt they did not have the time to implement the infrastructure required, but also that Heritrix and other existing web archiving technologies seemed like a ‘nuanced set of tools’ not fit for their purposes:

“There are 700 configuration lines. I don’t have time for this. It would be faster if I put a wget loop in a script and wrote down my IP address so that’s what I’m going to do. It was that level of—these tools don’t work because they were not designed for this use-case. And that to me was immediately and blatantly obvious. And it was just like no, this doesn’t do anything that we need it to do. What’s this thing do—it spits out a WARC record—what’s a WARC record? [...] Let’s record some stuff as JSON because that’s what everybody in the modern universe uses.” (Luke)

This passage reveals some of the challenges of the technical requirements of web archiving, in this case exacerbated by the urgency of the political climate. Luke went on to say that ‘the world [was] on fire’, and they, along with others, felt that there was ‘just no time to stop and have a discussion about best practices’. Whereas web archiving may in fact be a boundary object that brings together many different stakeholders who are accompanied by their own knowledge and expertise, there are still many barriers to entry. This observation is reminiscent of Star and Ruhleder’s remark that ‘one person’s standard is in fact another’s chaos’ (1996, p.378). In this case, interfacing with the tools and accepted standards of web archiving (e.g. WARCs) still require

a significant amount of *articulation work* (Strauss, 1988) to implement locally – or in other words, the work required to make things work. EDGI members (including those with technical skills) did not have immediate access to the tacit knowledge or ‘interpersonal and organizational networks’ (Star and Ruhleder, 1996, p.385) to help them successfully implement and operate these particular tools whilst also working in ‘crisis mode’. And as Augustine reflects, despite a feeling that EDGI and collaborators were aware of this gap in knowledge, they too felt that the urgency of the political moment nonetheless required action:

“I think we were aware—we were like ‘we don’t know how to use these tools in the right way’ [...]—that we were not doing things in the best way and we were probably re-creating something else. But I also think the moment was unique in a way that we shouldn’t have ignored and I’m glad we didn’t.”
(Augustine)

In the next section I summarise and reflect further on the political dimensions of EDGI’s collaborative efforts to archive environmental websites and data.

6.4 Reflections on the Politics and Impact of DataRescue

As of May 2017, DataRescue events nominated more than 63,076 seeds to the End of Term project and identified more than 22,000 datasets as ‘candidates for non-automated preservation’ (Walker et al., 2018) through their data archiving workflows. Their efforts, combined with the work of other contributors made the 2016 End of Term crawl the largest and most comprehensive crawl of the EPA.gov to date (Dillon et al., 2017; Lamdan, 2018). The impact of DataRescue can be visualised in a visualisation of End of Term 2016 nominations (Figure 6.8) that clearly demonstrates the significant contribution made by EDGI and Data Refuge volunteers in the submission of seeds to the project.

However, despite these successes several informants reflected on a certain degree of pushback from different communities about the guerrilla archiving/DataRescue efforts. In the aftermath of the project and speaking from the perspective of a local event organiser, Lamdan (2018, pp.236-238) outlines numerous shortcomings of the DataRescue efforts, arguing that: volunteers were not equipped with the sociotechnical expertise to archive and store .gov resources, the scale of the task and the public response were too great to manage the project effectively and in the end, it was difficult to sustain longterm commitment from volunteers. As further evidence of the futility of DataRescue, Lamdan (2018, p.239) points to the government-initiated snapshot of EPA.gov that was taken in January 2017, arguing that this snapshot provides the access inefficiently sought through community web archiving efforts.³¹

³¹<https://19january2017snapshot.epa.gov> (visited on 30th Jul. 2019)

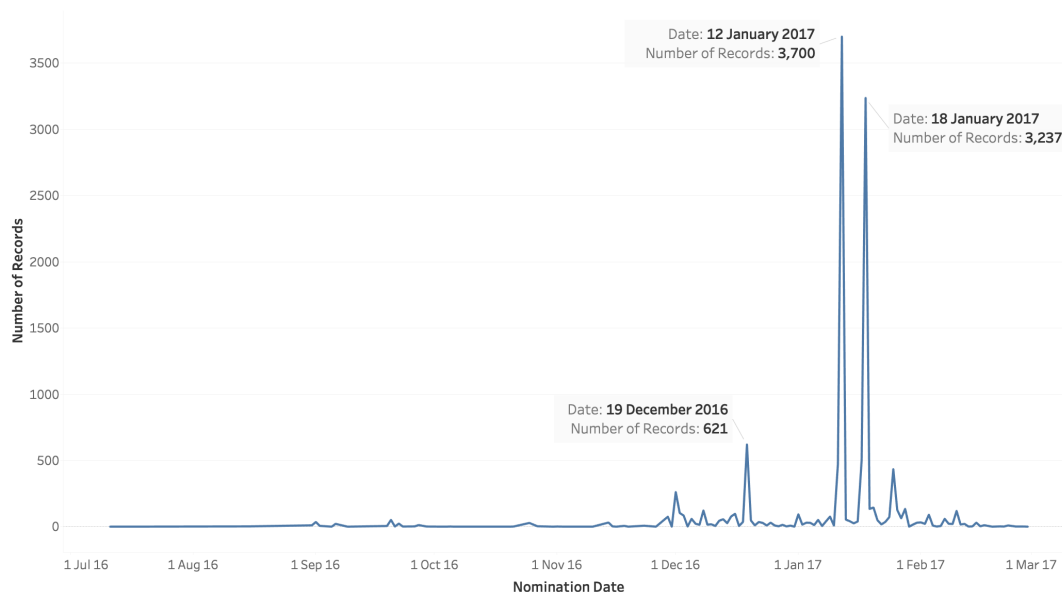


FIGURE 6.8: A graph of the 2016 End of Term nominations over time which shows the major peaks in nominations in approximate correlation with the initial DataRescue events in Toronto, Philadelphia and Indianapolis.³⁴

I raise these responses (and others below) to the DataRescue efforts to further emphasise the situated and subjective ways in which web archiving is positioned as a means for enacting data politics. And whilst I would generally support Lamdan's position that the problems of ensuring persistent public access to government data and information should be addressed through legislation, arguing that the EPA snapshot offers persistent public access misses many of the lessons of DataRescue, and works to further support the underlying argument of this thesis. In fact, the EPA snapshot is demonstrably incomplete. For example, it only includes the English language portions of the EPA site, excluding all Spanish content (espanol.epa.gov) and therefore further reducing historical access to EPA information on climate change for a substantial portion of the US public.³² Furthermore, the snapshot does not include numerous data portals that were once included on EPA.gov, nor what it considered 'large collections' of EPA.gov content due to disk size constraints.³³ In short, the snapshot is subject to the same subjective decision-making processes, resource constraints and sociotechnical contingencies of all efforts to archive the Web.

The subjective nature of practice and the perception of 'crisis' also extends to the web archiving community and works to further underscore the role and impact of technical expertise (and politics of boundary work) in the enactment of web archiving in practice. This observation is extended through a reflection amongst EDGI members that experts in web archiving believed the threat to environmental web resources

³²I was alerted to this point by the EDGI Web Monitoring team.

³³<https://19january2017snapshot.epa.gov/snapshot-help.html> (visited on 30th Jul. 2019)

³⁴The underlying data was exported from: <https://digital2.library.unt.edu/nomination/eth2016/reports/> (visited on 31st Jul. 2019)

that EDGI and others (e.g. Climate Mirror, Project Azimuth and Data Refuge) had identified was ‘overstated’:

“I would also say—this is how it has been put to me—is that a lot of people who did web archiving or digital preservation didn’t think that there was [...] as much of a threat and that it was overstated. And so looked at what was happening and were confused, but didn’t take that as an opportunity to be involved and help steer the direction of it—but instead saw that as a ‘oh they’re going to do this thing they don’t need to be’ and didn’t engage.”
(Augustine)

This comment from Augustine again highlights how different perspectives and forms of expertise frame how the threat was interpreted, and at the least, how an absence of a response in kind was perceived by some EDGI members. Several EDGI members were quick to point out though that it wasn’t that there was a lack of librarians and archivists involved in the wider movement around saving government data *per se* – Data Refuge was a collaboration with research data librarians and a number of DataRescue events were hosted by and with university libraries (for example, the Ann Arbor DataRescue was organised by the University of Michigan Libraries). Relatedly, as a volunteer consortium of the Internet Archive and other libraries, the End of Term project has written about the difficulties of managing the increased attention and offers of support they received during the 2016 crawling efforts:

“The community’s sudden desire to participate was unexpected, and the team struggled to find a way to harness all of this public energy in a productive way. Companies were interested in providing storage and computer infrastructure for the project. Individuals wanted to crawl content on their own and then contribute it to the project. People that didn’t know how they could help wanted to talk to the team about ways that they could contribute. The team was almost overwhelmed by eager assistants with nothing specific they could do” (Phillips and Phillips, 2019, p.29).

This point is extended by Jacobs and Bailey (2017) who described DataRescue and the associated media attention received by the EoT project as both a ‘blessing and a curse’, as they reportedly spent a significant amount of time fielding media interviews and explaining web archiving to journalists. In addition however, there was a distinct sense amongst some members from both inside and outside library/archival communities that, in general, challenges for collaboration stemmed from fundamental differences between the ways institutions and grassroots movements work to achieve social change. Speaking about the expertise of what they called the ‘larger data curation, data archiving, and federal information preservation field’, Gary frames these differences as a challenge for collaboration:

“[the way] I’ve framed this in my mind is we knew enough to get started but we didn’t know enough that those communities do, to know how such a Sisyphean task this is. And as I’m sure you’ve heard other people say, trying to work with these larger library organisations can be really challenging. Because I think that one of the reasons [...] is libraries are very often in the ‘for everybody and forever business’. Which makes it really hard to be nimble and try to respond to things. And there’s also a concern with being too responsive to short-term needs in libraries because then it just becomes a bunch of boutique projects or things that aren’t sustainable.” (Gary)

However, even more significant than the practical challenges of working across institutional and community boundaries, according to some, community-based web archiving also poses risks to the perception of objectivity in institutional recordkeeping.³⁵ The fear of partisanship can be seen in comments by Millar (2017) who warn against the potential negative consequences of collaborating with guerrilla archivists beyond the institutional boundaries of conventional memory work:

“The challenge with archivists reaching beyond our institutional boundaries, or even reaching beyond the scope of consulting and advising, is that we risk diminishing our credibility as objective recordkeepers and being seen as just as political as those in power” (Millar, 2017, p.69).

Despite the above attempts to reinforce the objectivity of recordkeeping and distance institutional involvement in the politics of guerrilla archiving, as others have detailed extensively (Brown and Davis-Brown, 1998; Zinn, 1977), here I assert that archiving and associated memory work is always a political project. *Web archiving as politics* deliberately positions web archives as intrinsically linked to both the contemporary politics of the object of collection and the ways that web archiving is negotiated and enacted through particular forms of expertise, value systems and practice. As the example of the emergence of EDGI and DataRescue have demonstrated, web archiving is being mobilised in an atmosphere of wide-spread public debates surrounding the value and defence of scientific knowledge-making. Through the mobilisation of over 1,500 people at approximately 50 events across the United States and Canada, DataRescue had an indelible impact on the shape of the archived Web at the transition to the Trump presidency. The intersection of stakeholders who mobilised to ‘rescue data’ highlighted the politics of expertise that occurs at the boundaries between web archiving practices and ultimately framed how the Web was archived in DataRescue and beyond.

³⁵An aversion to the politics associated with DataRescue was also seen to be reflected within the science community. One informant, Wesley, speculated about the difference between politics and partisanship and characterised an identifiable ‘reluctance’ amongst some in the science community to confront the politics of the moment for fear of being seen as partisan. There were others who warned against exaggerating the threat of the incoming administration, for example when Mark Serreze, a Director of a science data centre in Boulder, Colorado cautioned the scientific community at a scientific conference in December 2016 to ‘be careful not to overreact’ (Brennan, 2016).

EDGI's post-DataRescue work has been extensively informed by the above experiences of utilising and negotiating web archiving as a tool for government accountability and environmental justice. The Web Monitoring project that emerged in parallel to DataRescue, has sought to produce evidence-based reporting on 'meaningful changes' to the government Web (Rinberg et al., 2018). In collaboration with the Internet Archive, the Web Monitoring project continues to archive parts of the government Web associated with environmental policy and health and hazards, developing protocols and extensive reports on changes to web resources. The Data Together project emerged out of the sociotechnical challenges of DataRescue to advocate for community-based mechanisms for the stewardship and preservation of public data that are based in paradigms that support 'post-custodial stewardship' in collaboration with institutional and non-institutional partnerships (Abrams, 2017; Kelleher, 2017; Walker et al., 2018). Data Together therefore, fundamentally raises questions about how the Web is archived and who should control and steward the public data archives (Walker et al., 2018). These projects coupled with EDGI's emergent scholarly work on *environmental data justice* (Dillon et al., 2017, 2019; Walker et al., 2018) have all demonstrated the power of web archiving for shaping public discourses about data preservation in the face of a precarious Web.

6.5 Chapter Summary

The findings of this chapter offer some important insights into the research question: ***In what ways do web archival practices (the who, why and how) shape the archived Web?*** Through the lens of *web archiving as politics* I made the case that web archiving is inherently a political project. In the case of EDGI and DataRescue, politics both mobilised the use of web archiving as a political tool for ensuring access to public Web resources, and reflected various forms of politics (associated with expertise, values and decision-making) that occur at the intersecting social worlds engaged in web archiving. Here, the acts associated with deciding what to save and how are intrinsically linked, and worked to demonstrate the everyday politics of web archiving.

This chapter therefore argued that web archiving simultaneously *produces* and is *produced by* politics, an observation that fundamentally shapes the nature of who, why and how web archiving is done in practice. I began the chapter through a discussion of the emergence of the EDGI organisation in the wake of the US Presidential election of Donald Trump in November 2016. Here, I outlined the ways that the new administration (and previous precedents in Canada) helped shape the formation of the EDGI organisation and a subsequent DataRescue movement through the discourse of 'crisis'. I detailed the ways that the perception of the threat, urgency and uncertainty of the incoming administration's stance towards science data, and climate science (in particular) mobilised EDGI and others to preserve access to web-based environmental

data and information resources. Here, the political climate produced a movement of over 1,500 participants to engage with archiving the US government Web.

I then outlined a two-part argument that centred around the use of two STS concepts – boundary objects and boundary work – to first frame the ways that web archiving (as practices, artefacts and knowledge work) became a system of boundary objects that mobilised a diverse set of stakeholders around the mission to ‘rescue data’. Here I discussed a variety of ways that participants framed their motivations for enacting web archiving, including how web archiving was framed as a form of political action, enabling activists to ‘doing something’ in defence of public access to environmental data resources. And whilst EDGI and DataRescue signalled a sense of optimism for the collaborative potential of the expertise that participants brought to archiving the Web, boundary work highlighted the ways that this expertise played a part in shaping the nature of how practice was negotiated, challenged and enacted.

Through EDGI’s participation in the End of Term project and their own efforts to archive data in collaboration with DataRefuge, I discussed the creation of the DataRescue workflow, the End of Term nomination tool and the Archivers.space app. Each of these highlighted the ways that EDGI developed sociotechnical strategies for enacting web archiving and creating an inclusive community around DataRescue. The fundamental differences between how provenance was conceived and operationalised, for example, provided windows into the work that occurs at the boundaries of social worlds engaged in web archiving during DataRescue.

The general absence of volunteers and organisations with web archiving expertise in DataRescue (beyond the End of Term partners) was also discussed to underscore further barriers to entry for community members wishing to make use of web archiving in practice. I extended this discussion in the final section as a mechanism for reflecting on the impact and perceptions of DataRescue in the wider field whilst relating these to continuing debates around the supposed ‘objectivity’ of archival practice.

Whilst there are many other potentially fruitful avenues of discussion in EDGI’s work – both as they relate to web archiving as politics, as well as other facets of web archiving discussed in this thesis – the DataRescue movement neatly illustrates the co-constitution and significance of who, why and how web archiving is enacted. In the next and final chapter I examine the cumulative findings of this thesis and propose potential future work in this area.

7

Conclusions and Future Work

7.1 Facets of Web Archiving: A Review

In the preceding six chapters I have made the case for an examination of web archiving that centres an approach on the material and performative aspects of practice. In this final chapter I will first review my line of argument, methodology and findings from each empirical chapter before pivoting to a more general discussion that attempts to bring together this research. As a reminder, this thesis addressed the following research question:

In what ways do web archival practices (the who, why and how) shape the archived Web?

The initial three chapters of this thesis worked to frame the case for this research. In Chapter 1, I emphasised the growing role of web archives in the circulation of information and culture online by charting the ways that web archives are being used by academics, citizen activists, lawyers and journalists as resources for scholarly research and evidence-based accountability. Given these use cases, I outlined three fundamental concerns about the intrinsic connections between how the Web is archived, its future use, and our understandings of web archives, archivists and the Web. These concerns positioned the need for situating practice within the myriad of motivations, value-statements, mechanisms and circumstances under which web archiving is enacted. Chapter 2 charted the history of web archiving through a discussion of key initiatives, components and challenges encountered in web archiving and the use of the archived Web. I extended this discussion to problematise web archiving

from the point of view of archival theory and STS, by emphasising the materiality of practice and the critical need for research that considers and engages with the ways that web archiving is embedded within particular sets of sociotechnical relations that fundamentally shape how the Web is archived.

Given this case, in Chapter 3, I outlined an interpretive approach to researching web archiving that centred the materiality of practice. I positioned this research within a qualitative paradigm that drew on practice theory and multi-site ethnographic methods, before outlining my selection of the three sites of investigation. The choice of each site was driven by an initial framing of the field to position this thesis in contrast to existing research. I then discussed the data collection using ethnographic interviews, non/participant observation and documentary analysis across the three sites. I then frame the limitations of the approach as well as discuss the ethical implications of the methodological design. Subsequently I discuss the thematic analysis undertaken, as well as introduce the ways that my analysis borrowed from the other interpretive approaches to the use of metaphor, including the facet methodology approach (Mason, 2011).

The second half of this thesis presented the results of the empirical approach to the study of web archiving in three different sites: the Internet Archive, Archive Team and the Environmental Data & Governance Initiative (EDGI). The main observations from each site is summarised below.

Web Archiving as Infrastructure: Internet Archive

Chapter 4 was examined through the lens of *web archiving as infrastructure* at the Internet Archive. This approach enabled two interconnected observations about the ways that web archiving shapes the archived Web. The first is that web archiving is composed of a dynamic set of spatially and temporally located practices that shape the nature of what and how the Web is archived. This observation was demonstrated through my efforts to locate and contextualise the emergence of the Internet Archive within the 1990s/early 2000s in Silicon Valley. Here, I discussed the early goals of the Archive and its founder Brewster Kahle to build the online ‘Library of Alexandria 2.0’, in pursuit of particular sociotechnical imaginaries to enable ‘universal access to all knowledge’. I extended the view of this digital library to a description of the physical space in which the Archive occupies, in an effort to further situate the place in which local web archiving practices are shaped. In the round, the first part of this chapter worked to situate the ways that web archiving as infrastructure, and the practices this enables, emerges within particular social, cultural, economic, legal and ethical contexts that shape the material ways that web archiving is approached and enacted.

This observation is further supported by the analysis of the components of web archival labour through the second half of this chapter. Extending Downey’s (2014) concept of information labour, I outlined the case for four components of labour

(knowledge work, translation, maintenance and repair) that support the relational view of web archiving as infrastructure and emphasise the explicit ways that the everyday work of web archivists and engineers at the Internet Archive are shaping the archived Web. This analysis points towards a complex system of knowledge and maintenance work for prioritising which web assets to collect and repair. The Archive is leveraging their extensive existing archives for understanding networked linking behaviour in an effort to balance the breadth and depth of crawling activities, while discovering new sources for identifying websites to crawl based on measures of popularity, ‘novelty’ and sites that are endanger of going offline. The team has devised multiple mechanisms for identifying different types of ‘undesirable domains’, including rule-based link pattern-matching and the development of ‘gamified’ tools for the manual curation of sites.

Collectively, the efforts of the Archive can be seen as knowledge work, and these activities, seen in combination with other practices around the prioritisation, repair and maintenance of tools and archives all have ramifications for how web resources are transformed for use. It is the labour of non/human agents that enables the preservation and ingestion of information from the Web into the Archive, and then once again back to the Web where archives are reassembled via the Wayback Machine. Although imperfect, this labour is increasingly recognised as an essential element of the web architecture. The information labour and knowledge work of potential web archival users is therefore intimately tied to the web archival labour of the Internet Archive. As the global Wayback Machine currently provides access to billions of webpages – often inaccessible elsewhere – editorial decisions have implications for not only the fidelity of archived captures, but indeed whether or not certain parts of the Web are preserved at all.

Web Archiving as Culture: Archive Team

In Chapter 5, I examined *web archiving as culture* through the case of Archive Team, a community of ‘rogue’ web archivists. In this chapter I made two observations about web archiving. The first observation was that web archiving is produced within cultural worlds and particular systems of meaning. The second observation asserts that web archiving also creates and reinforces the cultural worlds they inhabit.

In the first half of the chapter I used ‘community’ as an analytical device through which to probe the cultural dimension of Archive Team’s web archiving practices. I first framed the origins of Archive Team through a discussion of how the collective emerged through the closure of AOL Hometown. I explored how Jason Scott and others have constructed a community around archiving the Web through the use of distributed technologies for communication and self-archiving. The community protocols for IRC and the wiki both enabled the work of Archive Team, but I argue they also signal an attentiveness to the broader sociotechnical values of Levy’s (2010)

'hacker ethic'. I then went on to describe two tenets of practice that I called *archival neutrality* and *brute force archiving* to describe the ways that Archive Team reflects an approach to 'radical web archiving'.

The second half of this chapter examined the web archiving practices of Archive Team through the case of archiving Tumblr 'Not Safe For Work' (NSFW) blogs in December 2018. I provided context to the project through a discussion of the circumstances that surrounded the NSFW ban on the Tumblr platform. Following this I provided an account of Archive Team's deployment of their Warrior project, as well as the groups use of media to recruit participants to the cause. The final section returned to the concepts of *archival neutrality* and *brute force archiving* to explore the ways that these tenets were engaged, challenged and contested through practice. The final sections reveal two further priorities, archival abundance and archival integrity that shaped the ways Archive Team members made decisions surrounding the removal and blacklisting of externally linked images, 'notes' (or re-blogs) and the use of login cookies. Participants regularly discussed and contested aspects of their collection practices, whilst overwhelmingly favouring approaches that enabled them to collect more, despite the risks of breaching the access protocols of sites like Tumblr.

Web Archiving as Politics: Environmental Data & Governance Initiative

In Chapter 6, I examined web archiving through the lens of *web archiving as politics*. I began by positioning web archiving as an inherently political project. Here I made two further observations about web archiving as practice. First, web archiving is mobilised by politics. In the case of EDGI and DataRescue, politics mobilised the use of web archiving as a tool for ensuring public access to the government Web. And second, web archiving produces politics through practice. Here, the project of DataRescue produced various forms of politics through the interaction of range of stakeholders in the process of enacting web archiving.

The chapter expanded on these observations to include a discussion of the emergence of the EDGI organisation in the wake of the election of Donald Trump as US President in 2016. Through the discourse of 'crisis' I detailed EDGI's reactions to the incoming administration as situated through three aspects of crisis, as defined by Rosenthal, Charles and 't Hart (1989): *threat*, *urgency* and *uncertainty*. I outlined the ways that the new administration was perceived as a threat to environmental data and scientific information pertaining to climate change, in particular. This coupled with previous precedents set by the Harper administration in Canada, prompted the formation of EDGI and the centring of web archiving as a tool for mitigating what they believed to be an incoming administration that was 'anti-science' and 'anti-environment'. I argue that this crisis discourse shaped their motivations for mobilising web archiving through the DataRescue movement, but also had implications for how they enacted web archiving in practice.

Through two regularly deployed STS concepts, *boundary objects* (Star and Griesemer, 2015) and *boundary work* (Gieryn, 1983) I then framed the ways that web archiving *produced* and was *produced by* politics. DataRescue was framed as a system of boundary objects that mobilised multiple publics to engage with web archiving during the c. 50 community-hosted events across North America. Through a discussion of the DataRescue workflow and tools used to seed the End of Term crawl and capture the ‘uncrawlable’, I discuss the politics that emerged through the boundary work, or practices that foreclose on different forms of knowledge production. In order to satisfy the diversity of needs and concerns brought by different actors, EDGI and the DataRescue project had to ‘translate’ these multiple perspectives to design sociotechnical interventions to achieve the collective goal of ‘rescuing data’. Consequently, despite the successes of the web archiving events, DataRescue also highlights the politics of web archiving as revealed through the challenges of boundary work that attempts to both translate and lay claim to particular types of expertise and work across organisational boundaries.

In conclusion, I reflected on the impact of DataRescue on the nature of what was archived as part of the End of Term crawl, before discussing some of the wider reactions to their efforts to archive environmental data. Here I emphasise the subjective nature of ‘crisis’ and from the perspective of informants, I frame the ways that some felt that the wider web archiving community did not engage with the activities of DataRescue, despite their expertise in this domain. I then reflect on the ways that some reactions can also be seen as attempts to project the value of objectivity and neutrality in archival and recordkeeping practices. I end the chapter by re-positioning and reasserting the ways that web archiving is inherently political.

7.2 Facets of Web Archiving: Thesis Contribution

In this section I would like to re-emphasise the strategic reasoning behind the choice of these-three-facets at these-three-sites, as a mechanism for exploring the wider applicability of the observations made through this empirical work. Through the course of this research, and in particular in the completion of field work, each of these facets became critical concepts for illustrating the transformative force of web archiving practices at each site. I initially drew on the approaches of Mason (2011) and Morgan (1997) and their use of metaphor as a mechanism for prioritising particular aspects of the research field in question. After an initial thematic analysis of the data from each site, I chose each facet (infrastructure, culture and politics) to illuminate the field in ways that offered the most evidence and explanatory power for how and why web archiving is done. In particular, the choice of these facets was also informed through an analytical comparison of web archiving at each of the three sites. By this I mean that my choice to emphasise culture at the Archive Team was not only made because it critically illuminated particular aspects of practice, but was also decided in relation

to the other sites and facets. Each of the facets were present in each of the sites of investigation and could have provided the lens through which to discuss the analysis of web archiving at each site. As described in Chapter 3, I chose to emphasise these particular pairings of sites and facets to provide critical windows into practices at the Internet Archive, Archive Team and EDGI.

Below, I demonstrate the portability and contribution of each facet (as critical concepts) through examples drawn from across the other sites of investigation. As a reminder, in each empirical chapter I framed the analysis through two overarching observations about the ways that each facet illuminated web archiving practices at each site. Here, using observations of practice from the Internet Archive, Archive Team and EDGI, I apply and provide evidence for the six facet observations in each of the other two sites.

Infrastructure

Web archiving is a view from somewhere, it is situated in space and time.

For Archive Team, the spatial and temporal arrangements of infrastructure both enabled and constrained the work of web archiving. For example, nearing the end of the Tumblr project participants frequently collaborated to figure out ways of extending the window of time during which they could download the NSFW blogs before they were removed by Tumblr. Here, Archive Team (as an internationally distributed community) collectively strategised the location of their own virtual machines (VMs) and server infrastructure in relation to the international location of Tumblr hosts and CDNs (content delivery networks). They scouted out ‘local mirrors’ both nearer to their own infrastructure but also within a timezone that would enable more time to crawl, with the added benefit of producing quicker, more efficient downloads (from being closer the source on the network). Here, the spatial and temporal arrangements of Archive Team’s infrastructure (coupled with their own situated technical expertise) facilitated web archiving in a way that enabled them to collect more blogs, at a faster rate and in a longer period of time than would have otherwise been possible. Furthermore, it also emphasises the point that although the Internet is global, speed and accessibility on the network are heavily subject to space and place.

Web archiving is enabled through heterogeneous forms of labour.

Each of the sites in this thesis have demonstrated the breadth of labour required to enable the collection and maintenance of web archives in practice. Here I take each component of web archival labour in turn, beginning with *knowledge work*. In the case of EDGI the entire DataRescue workflow and toolkit can be seen as a form of knowledge work. For example, the production of agency primers that worked

to comprehensively catalogue the US government webspace were a cornerstone of the seeding efforts for the End of Term crawls (Section 6.3.2). *Translation processes* can also be seen in EDGI's attempts to port the DataRescue workflow for archiving 'uncrawlable data' to the Archivists.space app (Section 6.3.2). In this case, informants described the subsequent tensions that arose when the workflow was translated into software that transformed both the nature of practice and revealed epistemological differences over the nature of provenance and the community's capacity to flexibly iterate on the workflow. In short, 'something was lost' in practice. Though not fully illustrated here *breakdown*, *maintenance* and *repair* can be seen through the entirety of EDGI's Web Monitoring project. EDGI has built an entire infrastructure around monitoring the federal web presence that includes a complex assemblage of people, protocols, crawlers, bots and server infrastructure. Various services regularly break down and require maintenance and repair to enable the work of web monitoring to proceed.

Culture

Web archiving is produced within cultural worlds that shape meaning through practice.

The case of EDGI and DataRescue also present an opportunity to reflect on the ways that culture is produced through what Swidler (2001, p.92) has called 'public ritual occasions' that establish new forms of social practice and associated meanings, and create opportunities for participants to align themselves with a community identity through participation.¹ Here, I transfer this concept to the case of DataRescue to argue that the election of Donald Trump created a 'disruption event' (Sewell, 1999) that forged a new community centred on visible, public displays of data activism. DataRescue promoted a set of practices through which an explicit 'culture of resistance' was performed through web archiving. The DataRescue events can then be seen as a series of 'public rituals of resistance', where participants established and asserted community membership, whilst aligning themselves with the values of the EDGI and DataRescue through practice. In summary, in the case of EDGI, the DataRescue events can be seen as a set of practices that communicated and enacted a culture through archiving the Web.

Web archiving creates and reinforces the cultural worlds they inhabit.

Through the case of the Internet Archive, I want to briefly explore the interlinking ways that the discourse of 'universal access to all knowledge' works to (re-)produce

¹Swidler (2001) uses the example of the original San Francisco Gay Freedom Day Parade to illustrate how visible, public displays of change or disruption (in this case, the public celebration of homosexuality) offer concrete and reoccurring mechanisms for communities to demonstrate their commitment to social causes, but also enrol new participants.

web archiving through practice. Here, web archiving practices in interaction with the discourse of ‘universal access to all knowledge’ enables an ‘observable object for the study of culture’ (Swidler, 2001, p.84) at the Internet Archive. Elsewhere, as Ben-David and Amram (2018) have recently framed, the system of web archiving practices that enable the Wayback Machine can also be seen as a form of epistemic practices that shape the production of ‘sociotechnical facts’. This framing, in combination with the concepts developed in this thesis surrounding the nature of practice as labour (Section 4.2), enables the view of Internet Archive practices as a form of epistemic culture (Knorr Cetina, 1999) that shapes the nature of archival production. This framing is complemented by work by Ivanov (2017) who uses practice theory to consider archival practices as a form of knowledge production that is shaped by particular cultural and organisational functions the archive performs. Epistemic culture can then be used to underscore the ways that the Internet Archive uses the results of the knowledge work practices identified in Section 4.3.1, to produce more knowledge in the form of more archives – a practice that I speculated can also be seen as a form of knowing capitalism (Thrift, 2005). Here, the epistemic practices of the Internet Archive – as varied and complex as they may be – work to create and reproduce the discourse of ‘universal access to all knowledge’ through everyday practice.

Politics

Web archiving is mobilised by politics.

As described in Chapter 4, the Internet Archive has regularly deployed web archiving at times of political transition. In collaboration with the Smithsonian and the Library of Congress, some of the Archive and Alexa Internet’s earliest targeted crawls were around the 1996 and 2000 US Presidential elections Section 4.2.1. During this time, the Archive crawled the websites of the Presidential candidates as a matter of recording the campaign agendas of those running. The 1996 collection was subsequently the first web archive exhibition on the Internet Archive (Figure 7.1), and the subject of an exhibit at the Smithsonian on Presidential websites (Kahle and Vaddillo, 2015). The 2000 archives became the subject of several early studies on the impact of Internet-based campaigning (e.g. Schneider and Foot, 2002). These early web archiving interventions both demonstrated the utility of web archiving for documenting the new age of web-based publication. But also, collaborating with high profile (and conventional) memory institutions like the Smithsonian and the Library of Congress gave the Internet Archive, in Kahle’s words, a sort of ‘blessing’ to move forward despite the legal uncertainties associated with web archiving at that time.

²<https://web.archive.org/web/19971011050034/http://www.archive.org:80/> (visited on 1st Aug. 2019)



FIGURE 7.1: The Internet Archive's web exhibition of the 1996 US Presidential election web archive.²

Web archiving produces politics.

In the case of archiving Tumblr, the politics produced by practice came in many forms throughout the research. Another form of politics can be seen in some of the reactions to archiving Tumblr not examined in Chapter 5. As discussed in Section 5.3.1, given the ways that the (pre-ban) Tumblr NSFW community evolved into a space dedicated to taboo topics and 'counterpublics' often considered fringe and marginalised 'in real life', some felt that the web archiving efforts of Archive Team were unfairly targeting the intimate content of this vulnerable online community. In practice, the 'politics of ethics' are observable through reactions to web archiving and Archive Team's subsequent responses. It was argued by others that transporting Tumblr NSFW posts to web archives therefore further marginalises the users they intended to help by divorcing content from the agency of their original creators (Figure 7.2).³ Whilst ostensibly fighting for the rights of Tumblr users, in this case, their right to consent to the de-contextualising of their content was outweighed by a mission to save, as one Archive Team member called it, a 'culturally significant portion of the Internet'.

Analytical Implications and Future Work

I want to propose potential avenues for future work that could both apply the approach taken here in other contexts, as well as extend this methodology to include

³This debate relates to work elsewhere on the nature of platform moderation intervention practices, as 'the entangled judgements of relevance, value and propriety' (Gillespie, 2018, p.196), and the 'social afterlife' of images when they become divorced from their original context when re-blogged on Tumblr (Tiidenberg, 2016, p.1566).

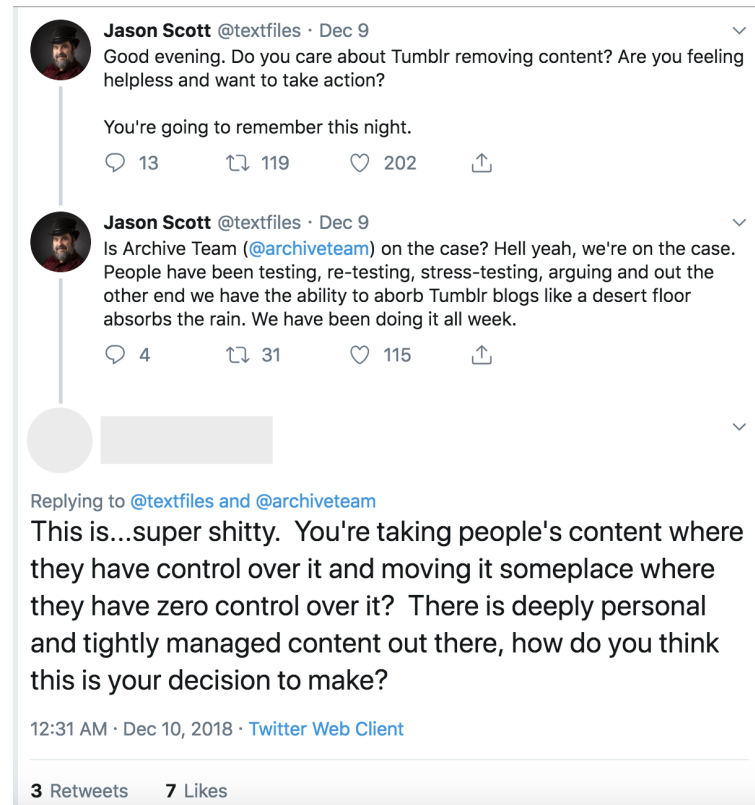


FIGURE 7.2: Screenshot of Twitter user reaction to Archive Team web archiving Tumblr NSFW.

additional facets for examining the sociomateriality of archival practices in a fruitful way. The first most obvious path is to apply these observations wholesale to each of the other sites considered in this thesis. As demonstrated above, there were already examples that point towards possible further insights using this empirical data. It would also be worth exploring the emergence of other observations within the chosen facets for each site. Whilst the facets presented here were the most persuasive, there will certainly be further insights into practice that can be gleaned from this empirical work.

In summary, this thesis centres the ways that web archiving – as infrastructure, culture and politics – shapes the nature of how the Web is archived. However, this is not to say that these are the only web archiving facets that could be observed. There will certainly be other facets of web archiving that deserve further attention in future work. One approach would be to explore these three facets and associated observations in other sites of web archiving. In particular, this work would greatly benefit from considering whether or not these observations hold true in the context of conventional memory institutions engaged in web archiving, like for example, national libraries and archives. Although I make no claims in this thesis as to whether or not the *specific practices* and approaches to web archiving observed at the Internet Archive, Archive Team and EDGI would have been observed elsewhere, my fundamental assertion is that *all web archiving can be seen as infrastructure, culture and*

politics.

Although the previous section is only a brief exercise in applying each of the facets to the other sites of investigation, it does offer further evidence for the applicability of these critical concepts across the Internet Archive, Archive Team and EDGI. These particular facets offer windows into web archiving practice that illuminate how web archiving is done and the ways in which different practices were influenced by various factors associated with organisational mission, values, community needs, technical provisions for web archiving and the dynamic negotiation of each in response to emergent obstacles and challenges. Each facet directly addresses the question of how web archiving shapes the archived Web by illustrating how these three organisations and community groups frame their own archival activities through different strategies that shape what parts of the Web are ‘saved’.

And despite their apparent differences, a common thread between the Internet Archive, Archive Team and EDGI can be seen in their networked capacity for experimentation and disruption in the web archiving field. These sites, in particular, highlight the ways that the emergence of the Web itself and the networked enrolment of new actors are disrupting expectations about the role of conventional memory institutions in the work of preserving cultural and scientific memory. Here, web archiving is representative of a particular form of memory work that has transitioned from being the sole charge of state-based actors or professionals within libraries and archives to what De Kosnik (2016) calls ‘rogue archiving’. This thesis therefore underscores the value of examining web archiving as it occurs outside of conventional memory institutions, national libraries and legal deposit. And in doing so, the site selection reflects a central component of sociological enquiry which is, according to Hughes (1971, p.53), ‘to understand how social values and collective arrangements are made and unmade: how things arise and how they change’. Examining the ways that web archiving practices are being enacted outside the ‘conventional’, also reveals the ways that the Web itself is being ‘made and remade’ in the context of these three sites of web archiving – a topic that I further explore below.

7.3 The Web They Want

To conclude this chapter I want to briefly reflect on how web archiving is mobilised by particular *sociotechnical imaginaries* (Jasanoff and Kim, 2009) that are intricately connected to the ways that archivists imagine the future of the Web and their role in it. To begin, as Finn (2018, p.6) documents, infrastructures are simultaneously ‘forward and backward facing’, where the idea of infrastructures-as-relational invites us to attend to both the ways that infrastructures are rooted in the foundations of particular standards and conventions, as well as the future-making possibilities that are enabled by their production and maintenance. Here, and as a jumping off point,

I argue that for many web archivists throughout this thesis, web archiving functions as a form of *infrastructural inversion* (Bowker and Star, 1999, pp.34-46) for the Web, or rather, web archiving becomes a collective apparatus for imagining and agitating for the future *Web they want*.

For the Internet Archive, the imaginary is motivated by the sociotechnical possibilities of ‘universal access to all knowledge’ which, as Chapter 4 supported, has positioned the Archive in the centre of a global network of web archiving. From its inception, the Archive has been rooted in Brewster Kahle’s vision for building the ‘Library of Alexandria 2.0’ which draws on several imaginaries associated with ‘the global brain’ and ‘smart machines’, or artificial intelligence technologies that, with the help of networks have the capacity to map and link the World’s knowledge. And when Kahle discussed these visions through reflections on the early motivations for the Internet Archive, Alexa Internet and the earlier WAIS system, the lines were often blurred between the ways that he simultaneously referred to the Web and his vision for the ‘global brain’ as one in the same. And whilst recognising that early metaphors for the Internet were often framed in a similar vein (e.g. ‘information superhighway’) – upon returning to Kahle’s imaginary, I can’t help but observe the ways that, at times – and with the Internet Archive, in particular – it is sometimes difficult to distinguish where the Web ends and the web archive begins.

In the first instance, the blurring of this place of overlap between the Web and the archive can be observed in some obvious places. The first is inside the web archive, where studies have illustrated the ways that the live Web seeps into the archive in the form of so-called ‘zombies’ and cobbled-together representations of the Web that *never was* (Ainsworth, Nelson and Sompel, 2015; Brunelle, 2012; Brunelle et al., 2015). However, with the ever-diversifying ways in which the Internet Archive is deliberately embedding itself into the Web architecture, this distinction between Web and archive is even more blurry. For example, as discussed briefly in Section 4.3.3, the Archive has in recent years been replacing all 404 links on the English language Wikipedia with links to snapshots in the Wayback Machine. As described elsewhere, the volunteer Wikimedia programmers employed ‘best guess’ solutions to finding the snapshot that best represented the intentions behind the original linking to external sources from within Wikipedia articles. Aside from obvious critiques around the heuristics for determining the intentionality behind linking practices, I use this example to extend beyond the remit of this thesis that addressed the ways that web archiving shapes the archived Web. Here I ask how is web archiving fundamentally changing the Web itself?

I propose that the ways in which imaginaries are imagined and materialised through practice are intimately tied to the ways that ‘the problem’ is conceived. Despite common framings of ephemerality as a technical fault of the Web’s architecture, this thesis has shown that conceptualisations of ephemerality should take into account the broader sociotechnical processes that shape the Web’s transience. Across these

sites of investigation web archivists have framed the problems of ephemerality (sometimes implicitly, sometimes explicitly) in ways that have focused on specific types of web resources (social media, science data), created and mobilised public(s) around particular forms of resistance, and informed strategies of direct action in practice.

Although steeped in a commitment to direct action, the long-view of the evolution of EDGI tells a story of reflexive archival practice; a shift from what they called ‘the sprint to marathon’. Given the perils of working ‘in crisis mode’ (Section ??) EDGI members regularly spoke of burn out, financial hardship and a general overcommitment. In this self-proclaimed move to ‘slow archiving’ EDGI has actively committed itself to tenets of working practices that incorporate principles of data justice, engage with questions of the potential for data harm and ongoing fears of re-producing the ‘single point of failure’ problem of centralised archives. In the case of EDGI, the imaginary – the Web they want – is a performative and continuous process that is dynamic, evolving. For EDGI what was once implicitly framed as a focus on collection (e.g. scraping and copying web data) and preservation, has evolved into an attentiveness to the wider issues of ‘single points of failure’ in the Web architecture. In collaboration, EDGI has shifted to exploring and enabling alternative ways for community-based stewardship of archives and data. For the imaginary, this may be characterised as the difference between imagining the Web as a place where state actors play a centralised role in the provision of access to public data to (perhaps, also) imagining a Web that isn’t burdened by centralised storage, but rather is enabled by networked communities hosting and stewarding the Web they want.

In addition to the focus on archiving environmental data and websites during DataRescue, EDGI members were also concerned with developing mechanisms to use web archives for government oversight. Here EDGI began the Web Monitoring project to use web archiving as a tool for tracking so-called ‘meaningful changes’ to the archived Web. Whilst the Web Monitoring project began to take shape, some EDGI members began to imagine the ways that the monitoring project could also be expanded to facilitate other areas of accountability in governance, beyond environmental science and policy. Here, the group envisioned an open platform that provided a ‘way to keep track of what governments do’. They called it *Everythingdotgov*.

In the months following the DataRescue SF Bay (at the University of California Berkeley) in February 2017, members penned a ‘one pager’ to describe their vision and circulated it amongst the EDGI members, prospective funders and partner organisations in order to mobilise support. Everythingdotgov emphasised the importance of providing a ‘public platform’ for facilitating access to federal data and knowledge as a ‘public resource’:

“The data and knowledge created and stored by the federal government is a public resource that belongs to the people. EDGI proposes a public platform – Everythingdotgov – that will archive government websites

and digital data and allow anyone to monitor how publicly available information changes over time. If governments hide or remove evidence that contradicts their interests, cut programs and regulations, or reverse progress towards open data, our platform will enable journalists and civil society organizations to hold them accountable” (Environmental Data and Governance Initiative, 2017).

In addition to building the platform to support their vision for Everythingdotgov, EDGI is now experimenting with ways to steward grassroots collaborative web archiving via the participation in the development and use of new protocols like Interplanetary File System (IPFS) to address the perceived need for distributed storage on the Web via the use of content-based addressing. The Data Together Project has emerged from this and presents an alternative imaginary for the Web – one that EDGI envisions as the manifestation of data justice principles to enable community-based infrastructure for stewarding data. In this web imaginary, access becomes thickened to move beyond the concept of ‘URL-accessible’, towards an access that is situated and steeped in community-based critique that fundamentally questions the power relations which underly data collection, stewardship and access. Here web archiving is envisioned as a mechanism for subverting mainstream/institutional narratives surrounding who stewards what and for whom.

The Data Together Project and IPFS is part of a wider network of practitioners and technologists advocating for the decentralised Web. And it is no coincidence that it was the inaugural *Decentralized Web Summit* that brought Vint Cerf (the co-inventor of TCP/IP) and Tim Berners-Lee (the inventor of the World Wide Web) to the Internet Archive – further signifying web archives as the site in which these future imaginaries for the Web will be shaped. The question remains, who will shape them?



Participant Information Sheet - Organisations

Participant Information Sheet – Organisation

Study Title: Observing Web Archival Practice

Researcher: Jessica Ogden

Ethics number: ERGO/FSHMS/23189

Please read this information carefully before deciding to take part in this research. If you are happy to participate you will be asked to sign a consent form.

What is the research about?

This study is being undertaken as part of a PhD research project on web archival practice. Much of the focus of the web archiving community has been on the continued development of technologies and practices for web collection development, with increased attention in recent years on facilitating the scholarly use of web archives. This research will take a step back to consider the place of web archives in light of postmodernism, 'the archival turn' and emergent questions over the ever-expansive role of memory practices and the archive in everyday life. The mechanisms and circumstances surrounding the production of web archives are fundamental to understanding them as 'new forms of social data'. This research proposes to re-situate web archives as places of knowledge and cultural production in their own right, by implicating both the web archivist and technologies in the shaping of the 'politics of ephemerality' that lead to the creation, maintenance and use of web archives. In short: **How does web archival practice (the who, what and how) shape what is known about the Web?**

This study aims to address the following objectives:

- 1) To identify key underlying assumptions about what the Web is, what of the contemporary Web is (or isn't) being archived, and the relative affordances for web archival practice and scholarly use.
- 2) To consider the *performativity of web archiving*, and the ways in which the practices of selection, collection and classification are forms of *knowledge production*, and thus shape what is known about the Web.
- 3) To examine the implications for a socio-technical understanding of web archives - by observing the interplay of both social and technological agents in the production of web memory practice(s).

This will be achieved through the use of ethnographic methods to produce an in-depth understanding of the assumptions, motivations, and technologies that inform practice(s) within different communities engaged with archiving the Web. The research will be carried out within three to four distinct 'communities of practice' (pending consent) – some of which involve formally dedicated organisations engaged in web archiving such as libraries, archives and other non-profit groups, as well as other less formal, distributed collectives of volunteers and hobbyists. The overall aim is to take into account and document the wider social, cultural, economic and technical landscape in which each community of practice resides.

This study is supported by the Web Science Centre for Doctorial Training at the University of Southampton (UK).

Why have I been chosen?

You have been approached because you have been identified as an organisation or member of an organisation that was or is involved in web archiving in some way – including but not limited to activities associated with web harvesting and indexing, metadata or description practices, supporting research, developing archival policies, services, tools and standards, etc.

What will happen if I take part?

The study is taking an ethnographic approach, which includes a flexible period of on-location **observation** with employees, team members and volunteers involved in web archiving at your organisation. The researcher will observe (as a participant or non-participant, as appropriate) the everyday activities of web archivists and other individuals that support and engage with web archiving. This will involve allowing the researcher to be present at the organisation and to shadow employees and team members in the office during their everyday work activities. The researcher will use digital photography to record some observations and may follow-up with unstructured engagements with individuals, including asking informal questions and clarifications on aspects of activities, as observed.

All observations will be contingent on prior organisational consent, where individuals will be given the opportunity to opt-out should they not wish to take part in the observations.

Walking interviews will be used to contextualise the observations by enabling individual participants to describe their working experience as it exists within the actual space and place of the organisation. Individual participants will be asked to describe and explain their work as it takes place in different physical (and digital) spaces associated with archiving. Occasional follow-up questions will be asked to clarify aspects of each tour, and may also include the use of 'screen casts' and other digital recording techniques (e.g. photography) to record the tours of working environments. Each walking interview is expected to last no longer than 1 hour.

Biographical interviews will be used to further situate the observations and give a narrative element to understanding the historical development of web archival practice within your organisation or community. This will involve an in-depth semi-structured interview with the investigator, focusing on the participant's involvement in web archiving and lasting approximately 1.5-2 hours.

All interviews will be contingent on prior individual consent, and will be audio recorded. Some individuals may be recruited for one or both types of interviews, but individuals are under no obligation to participate in either.

Where available, the researcher will request access to any archival policy and practice documentation produced by your organisation.

Are there any benefits in my taking part?

It is expected that the study will add to current knowledge about the state and development of web archival practice and directly contribute to the investigator's graduate-level research. Furthermore (and where applicable), it is the researcher's hope that the study may also incentivise and promote further interaction with web archives within the organisation and the wider community of potential users of web archives.

Are there any risks involved?

The expected risks to involvement are considered minimal, and primarily associated with the 'pseudonymisation' of observations. All individual participant names will be given pseudonyms in the reporting of the findings of this research, though informants should be aware that full anonymity may not be possible in the context of the relatively small community of individuals and organisations associated with web archiving. The inclusion of organisation names and affiliations in the reporting will be negotiated on a case by case basis. In the event that individuals or organisations do not wish their affiliation to be revealed, organisations will be given a pseudonym, as well.

Will my participation be confidential?

Although this study will not collect personal or sensitive data, anything recorded during the context of the observations deemed confidential will remain so.

What happens if I change my mind?

Organisations and individuals may withdraw at any time and for any reason, at which point your data will be removed from the study.

What happens if something goes wrong?

Should you have any concern or complaint, please contact the University of Southampton Head of Research Governance (+44 2380 595058, rgoinfo@soton.ac.uk).

Where can I get more information?

Should you require any further information on the study, please contact the investigator directly (Jessica Ogden: jessica.ogden@soton.ac.uk).



Participant Information Sheet - Online
Community

Participant Information Sheet – Online Community

Study Title: Observing Web Archival Practice

Researcher: Jessica Ogden

Ethics number: ERGO/FSHMS/23189

Please read this information carefully before deciding to take part in this research. If you are happy to participate you will be asked to sign a consent form.

What is the research about?

This study is being undertaken as part of a PhD research project on web archival practice. Much of the focus of the web archiving community has been on the continued development of technologies and practices for web collection development, with increased attention in recent years on facilitating the scholarly use of web archives. This research will take a step back to consider the place of web archives in light of postmodernism, 'the archival turn' and emergent questions over the ever-expansive role of memory practices and the archive in everyday life. The mechanisms and circumstances surrounding the production of web archives are fundamental to understanding them as 'new forms of social data'. This research proposes to re-situate web archives as places of knowledge and cultural production in their own right, by implicating both the web archivist and technologies in the shaping of the 'politics of ephemerality' that lead to the creation, maintenance and use of web archives. In short: **How does web archival practice (the who, what and how) shape what is known about the Web?**

This study aims to address the following objectives:

- 1) To identify key underlying assumptions about what the Web is, what of the contemporary Web is (or isn't) being archived, and the relative affordances for web archival practice and scholarly use.
- 2) To consider the *performativity of web archiving*, and the ways in which the practices of selection, collection and classification are forms of *knowledge production*, and thus shape what is known about the Web.
- 3) To examine the implications for a socio-technical understanding of web archives - by observing the interplay of both social and technological agents in the production of web memory practice(s).

This study is supported by the Web Science Centre for Doctorial Training at the University of Southampton (UK).

Why have I been chosen?

You have been approached because you have been identified as a person who was or is involved in web archiving in some way – including but not limited to activities associated with web harvesting and indexing, metadata or description practices, supporting research, developing archival policies, services, tools and standards, etc.

What will happen to me if I take part?

The study is using **biographical interviews** to investigate web archiving within different communities of practice. The interviews will involve an in-depth semi-structured interview with the investigator, focusing on your involvement in web archiving and lasting approximately 1.5-2 hours. The interviews are designed to give explore the underlying motivations and intentions driving web archiving, as well as provide a narrative element to understanding the

historical development of archiving within your community or organisation. All interviews will be digitally recorded either via Skype or an audio recorder in face-to-face interviews.

Are there any benefits in my taking part?

It is expected that the study will add to current knowledge about the state and development of web archival practice and directly contribute to the investigator's graduate-level research.

Are there any risks involved?

The expected risks to involvement are considered minimal, and primarily associated with the 'pseudonymisation' of observations. All individual participant names will be given pseudonyms in the reporting of the findings of this research, though informants should be aware that full anonymity may not be possible in the context of the relatively small community of individuals and organisations associated with web archiving. The inclusion of organisation/community names and affiliations in the reporting will be negotiated on a case by case basis. In the event that individuals or organisations do not wish their affiliation to be revealed, organisations will be given a pseudonym, as well.

Will my participation be confidential?

Although this study will not collect personal or sensitive data, any interview data deemed confidential will remain so.

What happens if I change my mind?

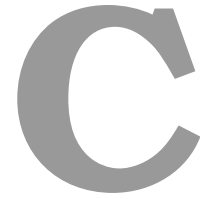
You may withdraw at any time and for any reason, at which point your data will be removed from the study.

What happens if something goes wrong?

Should you have any concern or complaint, please contact the University of Southampton Head of Research Governance (+44 2380 595058, rgoinfo@soton.ac.uk).

Where can I get more information?

Should you require any further information on the study, please contact the investigator directly (Jessica Ogden: jessica.ogden@soton.ac.uk).



Participant Information Sheet - EDGI

Participant Information Sheet – Organisation (EDGI)

Study Title: Observing Web Archival Practice

Researcher: Jessica Ogden

Ethics number: ERGO/FSHMS/23189

Please read this information carefully before deciding to take part in this research. If you are happy to participate you will be asked to sign a consent form.

What is the research about?

This study is being undertaken as part of a PhD research project which considers the mechanisms and circumstances surrounding the production of web archives. The overall aim of the project is to understand the motivations underlying different web archiving initiatives, and the ways in which these are transformed into practices, policies and technologies that shape the nature of information preservation and access. In short: **How does web archival practice (the who, what and how) shape what is known about the Web?** The research will be carried out within 3-4 distinct ‘communities of practice’ (pending consent). Further framing of the research and methods can be found in a recently published paper by the researcher: <http://dl.acm.org/citation.cfm?doid=3091478.3091506> or on the researcher’s website: <https://archivingtheweb.me/>. This study is supported by the Web Science Centre for Doctorial Training at the University of Southampton (UK).

Why have I been chosen?

You have been approached because you have been identified as an organisation or member of an organisation that was or is involved in web archiving in some way – including but not limited to activities associated with web harvesting and indexing, metadata or description practices, supporting research, developing archival policies, services, tools and standards, etc.

What will happen if I take part?

Online/offline participant observation within EDGI public/member spaces will be used to engage with the everyday practices of the community. Observations will be recorded whilst participating in different teams, with a particular focus on working with the *Archiving/Web Monitoring Project*, *Data Together* and any activities that support of the collection and storage of archived data. The researcher will record reflexive observations describing *what is done, made/used and said* (about web archiving practices) and may involve in-situ follow-ups with individuals, including asking informal questions and clarifications on aspects of activities, as observed. Participants will not be quoted in observations without explicit individual consent. Observations will require presence and access to organisational online spaces (including Slack channels, online meeting spaces, and other communication channels), as appropriate. All observations will be contingent on prior organisational consent, where individuals will be given the opportunity to opt-out should they not wish to take part in the observations.

Walk-throughs will be used to contextualise the observations by enabling individual participants (including the researcher) to describe their working experience in real-time. Participants will be asked to describe and explain their work as it takes place in different physical (and digital) spaces associated with activities in question. Occasional follow-up questions will be asked to clarify aspects of the walk-through, and may also include the use of ‘screen-casts’ and other digital recording techniques (e.g. photography) to record working environments. Each walking interview is expected to last no longer than 1 hour.

Ethnographic interviews will be used to further situate the observations and give a narrative element to understanding both the historical and contemporary development of web archival

practices within EDGI. This will involve in-depth semi-structured interviews with the researcher, focusing on the member's involvement in EDGI and lasting approximately 1-2 hours, as appropriate. All interviews will be contingent on prior individual consent, and will be audio recorded. Some individuals may be recruited for one or both types of interviews, but individuals are under no obligation to participate in either.

Documentary evidence will be used to further contextualise community (web archiving) activities, including for example, the use of reports, blog posts, meeting notes, previously recorded meetings, project documentation, media coverage and press releases, Github commits and comments. Any materials deemed inappropriate for the study will not be used.

Are there any benefits in my taking part?

Through the use of interviews, the study could also be used to document the community's 'institutional memory', providing both a retrospective (through oral history accounts and interviews) and a contemporary snapshot of the history of EDGI for future use. It is expected that the study will add to current knowledge about the state and development of web archival practice and directly contribute to the investigator's graduate-level research. Furthermore (and where applicable), it is the researcher's hope that the study may also incentivise and promote further interaction with web archives within the organisation and the wider community of potential users of web archives.

Are there any risks involved?

The expected risks to involvement are considered minimal, and primarily associated with the 'pseudonymisation' of observations and findings. All individual participant names will be given pseudonyms in the reporting of the findings of this research, though informants should be aware that full anonymity may not be possible in the context of the relatively small community of individuals and organisations associated with web archiving. The inclusion of organisation names and affiliations in the reporting will be negotiated on a case by case basis. In the event that individuals or organisations do not wish their affiliation to be revealed, organisations will be given a pseudonym, as well. All off-the-record spaces will be treated as confidential, including Slack channels and information deemed confidential during one-on-one interactions and interviews with the researcher. This study will not make records of information or data deemed sensitive to participants.

Will my participation be confidential?

Although this study will not collect personal or sensitive data, anything recorded during the context of the observations deemed confidential will remain so.

What happens if I change my mind?

Organisations and individuals may withdraw at any time and for any reason, at which point your data will be removed from the study.

What happens if something goes wrong?

Should you have any concern or complaint, please contact the University of Southampton Head of Research Governance (+44 2380 595058, rgoinfo@soton.ac.uk).

Where can I get more information?

Should you require any further information on the study, please contact the investigator directly (Jessica Ogden: jessica.ogden@soton.ac.uk).



Consent Form - Organisations

CONSENT FORM – Organisation

Study title: Observing Web Archival Practice

Researcher name: Jessica Ogden

Ethics reference: ERGO/FSHMS/23189

Please initial the box(es) if you agree with the statement(s):

I have read and understood the information sheet ([12 August 2016 – 23189_information_v3_AppA](#)) and have had the opportunity to ask questions.

☐

I consent to the organisation and members of our organisation taking part in this study and agree for their interactions to be observed, recorded and used for the purpose of this study.

☐

I understand that individuals will be given pseudonyms in reports of the research but that full anonymity may not be guaranteed. I understand that organisational names will be maintained unless otherwise agreed.

☐

I understand participation in this study is voluntary and that the organisation may withdraw at any time without their legal rights being affected.

☐

Data Protection

I understand that information collected about me during my participation in this study will be stored on a password protected computer and that this information will only be used for the purpose of this study.

Name of Organisation (print name)

Name of Line Manager/Director (print name)

Signature of Line Manager/Director

Date



Consent Form - Individuals

CONSENT FORM - Individual

Study title: Observing Web Archival Practice

Researcher name: Jessica Ogden

Ethics reference: ERGO/FSHMS/23189

Please initial the box(es) if you agree with the statement(s):

I have read and understood the information sheet ([12 August 2016 – 23189_information_v3_AppA](#) or AppB, as appropriate) and have had the opportunity to ask questions about the study.

☐

I agree to take part in this study.

☐

I consent to the interview being recorded.

☐

I consent to anonymised quotes from the interview being used in publications arising out of this study.

☐

I understand my participation is voluntary and that I may withdraw at any time without my legal rights being affected.

☐

Data Protection

I understand that information collected about me during my participation in this study will be stored on a password protected computer and that this information will only be used for the purpose of this study.

Name of participant (print name).....

Signature of participant.....

Date.....

Bibliography

- Abbate, Jane (1999). *Inventing the Internet*. Cambridge, MA; London, England: MIT Press.
- Abrams, Stephen (2017). 'Curation is Not a Place: Post-Custodial Stewardship for a Do-It-Yourself World'. Conference Presentation. DLF 2017 Forum. Pittsburgh, Pennsylvania. URL: <https://escholarship.org/uc/item/0wd3f1x4> (visited on 31st July 2019).
- Abu-Lughod, Lila (1990). 'The Romance of Resistance: Tracing Transformations of Power Through Bedouin Women'. In: *American Ethnologist* 17.1, pp. 41–55. ISSN: 0094-0496. URL: <https://www.jstor.org/stable/645251> (visited on 25th May 2019).
- Ainsworth, Scott G., Michael L. Nelson and Herbert van de Sompel (2015). 'Only One Out of Five Archived Web Pages Existed As Presented'. In: *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. HT '15. New York, NY, USA: ACM, pp. 257–266. URL: <http://doi.acm.org/10.1145/2700171.2791044> (visited on 11th Jan. 2017).
- Allen, Laurie (2017). *Data Refuge Rests on a Clear Chain of Custody*. PPEH Lab. URL: <http://web.archive.org/web/20170223160409/https://www.ppehlab.org/blogposts/2017/2/1/data-refuge-rests-on-a-clear-chain-of-custody> (visited on 23rd May 2019).
- Allen, Laurie et al. (2017). 'Building Data Refuge: From Bucket Brigade to Sustainable Action'. Conference. Coalition for Networked Information Spring 2017 Membership Meeting. Albuquerque, New Mexico, USA. URL: <https://youtu.be/UgEpHsMJgZA> (visited on 22nd May 2019).
- AlSum, Ahmed et al. (2014). 'Profiling web archive coverage for top-level domain and content language'. In: *International Journal on Digital Libraries* 14.3, pp. 149–166. URL: <http://dx.doi.org/10.1007/s00799-014-0118-y> (visited on 25th Mar. 2016).
- Alvesson, Mats and Dan Karreman (2000). 'Varieties of Discourse: On the Study of Organizations through Discourse Analysis'. In: *Human Relations* 53.9, pp. 1125–1149. URL: <https://doi.org/10.1177/0018726700539002>.
- Anderson, Kimberly D. (2011). 'Appraisal Learning Networks: How University Archivists Learn to Appraise Through Social Interaction'. PhD. Los Angeles: University

- of California. URL: http://works.bepress.com/kimberly_anderson/9/ (visited on 13th Jan. 2016).
- Anderson, R. J. (1994). 'Representations and requirements: the value of ethnography in system design'. In: *Human-Computer Interaction* 9.3, pp. 151–182. DOI: 10.1207/s15327051hci0902_1. URL: http://www.tandfonline.com/doi/abs/10.1207/s15327051hci0902_1 (visited on 25th Sept. 2016).
- Ankerson, Megan Sapnar (2012). 'Writing web histories with an eye on the analog past'. In: *New Media & Society* 14.3, pp. 384–400. URL: <https://doi.org/10.1177/1461444811414834>.
- (2015). 'Read/Write the Digital Archive: Strategies for Historical Web Research'. In: *Digital Research Confidential: The Secrets of Studying Behavior Online*. Ed. by Eszter Hargittai and Christian Sandvig. Cambridge and London: MIT Press, pp. 29–54.
- (2018). *Dot-Com Design: The Rise of the Usable, Social, Commerical Web*. New York: New York University Press.
- Archive-It (n.d.). *Collecting Organizations*. Archive-It.org. URL: https://archive-it.org/explore?falpha=f_organizationType:true (visited on 10th Jan. 2017).
- Archive Team (2016). *Vine*. Wiki. URL: <http://www.archiveteam.org/index.php?title=Vine> (visited on 7th Jan. 2017).
- (n.d.[a]). *ArchiveTeam Warrior*. Wiki. URL: http://archiveteam.org/index.php?title=ArchiveTeam_Warrior (visited on 12th Jan. 2017).
- (n.d.[b]). *Dev/Infrastructure*. Wiki. URL: <https://www.archiveteam.org/index.php?title=Dev/Infrastructure> (visited on 13th Feb. 2019).
- Armstrong, Chad (2011). *Racks of petabytes of data*. Flickr. URL: https://live.staticflickr.com/6093/6224869899_f1a5989216_b.jpg (visited on 27th July 2019).
- Arnold, Hillel (2016). *Critical Work: Archivists as Maintainers*. hillelarnold.com. URL: <http://hillelarnold.com/blog/2016/08/critical-work/> (visited on 30th Sept. 2017).
- Arvidson, Allan and Frans Lettenström (1998). 'The Kulturarw3 Project - The Swedish Royal Web Archive'. In: *The Electronic Library* 16.2, pp. 105–108. URL: <http://dx.doi.org/10.1108/eb045623>.
- Bailey, Jefferson (2013). 'Disrespect des Fonds: Rethinking Arrangement and Description in Born-Digital Archives'. In: *Archive Journal* 3. URL: <http://www.archivejournal.net/issue/3/archives-remixed/disrespect-des-fonds-rethinking-arrangement-and-description-in-born-digital-archives/> (visited on 9th Mar. 2015).
- (2016). *10 Years of Archiving the Web Together*. Internet Archive Blogs. URL: <http://blog.archive.org/2016/10/25/10-years-of-archiving-the-web-together/> (visited on 8th Jan. 2017).
- Bailey, Jefferson et al. (2014). *Web Archiving in the United States: A 2013 Survey*. The National Digital Stewardship Alliance, pp. 1–24. URL: http://www.digitalpreservation.gov/documents/NDSA_USWebArchivingSurvey_2013.pdf (visited on 28th Jan. 2016).

- Bailey, Jefferson et al. (2017). *Web Archiving in the United States: A 2016 Survey*. The National Digital Stewardship Alliance, pp. 1–32. URL: http://ndsa.org/documents/WebArchivingintheUnitedStates_A2016Survey.pdf (visited on 15th Mar. 2017).
- Ballard, Mark (2013). *Conservatives erase Internet history*. Computer Weekly. URL: <https://www.computerweekly.com/blog/Public-Sector-IT/Conservatives-erase-Internet-history> (visited on 13th Mar. 2019).
- Barad, Karen (2003). 'Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter'. In: *Signs* 28.3, pp. 801–831. URL: <http://www.jstor.org/stable/10.1086/345321>.
- Barbour, Rosaline S. (1998). 'Mixing Qualitative Methods: Quality Assurance or Qualitative Quagmire?' In: *Qualitative Health Research* 8.3, pp. 352–361. URL: <https://doi.org/10.1177/104973239800800306>.
- Barry, Rick (2010). 'Opinion piece - electronic records: now and then'. In: *Records Management Journal* 20.2, pp. 157–171. URL: <http://dx.doi.org/10.1108/09565691011064304>.
- Battley, Belinda (2013). 'Finding aids in context: using Records Continuum and Diffusion of Innovations models to interpret descriptive choices'. In: *Archives and Manuscripts* 41.2, pp. 129–145. URL: <http://dx.doi.org/10.1080/01576895.2013.793164>.
- Baxter, Pamela and S. Jack (2008). 'Qualitative Case Study Methodology: Study Design and Implementation for Novice Researchers'. In: *The Qualitative Report* 13.4, pp. 544–559. URL: <http://nsuworks.nova.edu/tqr/vol13/iss4/2> (visited on 25th Sept. 2016).
- Beer, David and Mark Taylor (2013). 'The Hidden Dimensions of the Musical Field and the Potential of the New Social Data'. In: *Sociological Research Online* 18.2, p. 14. URL: <http://www.socresonline.org.uk/18/2/14.html> (visited on 9th Apr. 2016).
- Ben-David, Anat (2016). 'What does the Web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain'. In: *New Media & Society* 18.7, pp. 1103–1119. URL: <http://dx.doi.org/10.1177/1461444816643790>.
- Ben-David, Anat and Adam Amram (2018). 'The Internet Archive and the sociotechnical construction of historical facts'. In: *Internet Histories* 2.1-2, pp. 179–201. URL: <https://doi.org/10.1080/24701475.2018.1455412>.
- Ben-David, Anat and Hugo Huurdeman (2014). 'Web Archive Search as Research: Methodological and Theoretical Implications'. In: *Alexandria* 25.1, pp. 93–111.
- Bergman, Rachel et al. (2019). *Erasing the Affordable Care Act: Using Government Web Censorship to Undermine the Law*. Trend Report. Washington, D.C.: Sunlight Foundation's Web Integrity Project. URL: <http://sunlightfoundation.com/wp-content/uploads/2019/05/Erasing-the-ACA-Using-Web-Censorship.pdf> (visited on 12th July 2019).

- Berners-Lee, Tim, Mark Fischetti and Michael L Dertouzos (2000). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. Harper Information. ISBN: 0-06-251587-X.
- Blok, Anton (1972). 'The Peasant and the Brigand: Social Banditry Reconsidered'. In: *Comparative Studies in Society and History* 14.4, pp. 494–503. ISSN: 1475-2999, 0010-4175. DOI: 10.1017/S0010417500006824. URL: <https://doi.org/10.1017/S0010417500006824> (visited on 28th Jan. 2019).
- Boellstorff, Tom (2012). 'Rethinking Digital Anthropology'. In: *Digital Anthropology*. Ed. by Heather A. Horst and Daniel Miller. London: Bloomsbury Publishing, pp. 39–60.
- Boin, Arjen and Paul 't Hart (2007). 'The Crisis Approach'. In: *Handbook of Disaster Research*. Ed. by Havidán Rodríguez, Enrico L. Quarantelli and Russell R. Dynes. New York, NY: Springer, pp. 42–54.
- Borman, Kathryn M., Margaret D. LeCompte and Judith Preissle Goetz (1986). 'Ethnographic and Qualitative Research Design and Why It Doesn't Work'. In: *American Behavioral Scientist* 30.1, pp. 42–57.
- Bourdieu, Pierre (1977). *Outline of a Theory of Practice*. Trans. by Richard Nice. English Language Edition. Cambridge: Cambridge University Press.
- (1991). *Language and Symbolic Power*. Trans. by Gino Raymond and Matthew Adamson. Cambridge, MA: Harvard University Press.
- Bowker, Geoffrey C. (2005). *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.
- Bowker, Geoffrey C. and Susan Leigh Star (1999). *Sorting Things Out: Classification and Its Consequences*. Paperback. Boston, MA: MIT Press.
- Bowker, Geoffrey C. et al. (2010). 'Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment'. In: *International Handbook of Internet Research*. Ed. by Jeremy Hunsinger, Lisbeth Klastrop and Matthew Allen. Dordrecht: Springer Netherlands, pp. 97–117. ISBN: 978-1-4020-9789-8. URL: https://doi.org/10.1007/978-1-4020-9789-8_5.
- Bowman, Lisa M. (2002). *Net archive silences Scientology critic*. CNet.com. URL: <https://web.archive.org/web/20021004040612/http://news.com.com/2100-1023-959236.html> (visited on 28th July 2019).
- boyd, danah m. and Kate Crawford (2012). 'Critical Questions for Big Data'. In: *Information, Communication & Society* 15.5, pp. 662–679.
- Bragg, Molly and Kristine Hanna (2013). *The Web Archiving Life Cycle Model*. Tech. rep. The Archive-It Team and Internet Archive. URL: http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf (visited on 1st Mar. 2016).
- Braun, Virginia and Victoria Clarke (2006). 'Using thematic analysis in psychology'. In: *Qualitative Research in Psychology* 3.2, pp. 77–101. ISSN: 1478-0887. DOI: 10.1191/1478088706qp063oa. URL: <https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa> (visited on 20th July 2019).

- Brennan, Charlie (2016). 'Boulder scientists question threat to climate data under Trump, caution against overreaction'. In: *The Daily Camera*. URL: http://www.dailycamera.com/science_environment/ci_30660844/boulder-scientists-question-threat-climate-data-under-trump (visited on 10th Apr. 2019).
- British Library Web Archiving Team (2014). *The British Library Collection Development Policy for websites*. URL: http://www.bl.uk/aboutus/stratpolprog/digi/webarch/bl_collection_development_policy_v3-0.pdf (visited on 20th Mar. 2016).
- Brown, Adrian (2006). *Archiving Websites: A Practical Guide for Information Management Professionals*. London: Facet.
- Brown, Richard Harvey and Beth Davis-Brown (1998). 'The making of memory: the politics of archives, libraries and museums in the construction of national consciousness'. In: *History of the Human Sciences* 11.4, pp. 17–32.
- Brügger, Niels (2008). 'The Archived Website and Website Philology'. In: *Nordicom Review* 29.2, pp. 155–175. URL: http://www.nordicom.gu.se/sites/default/files/kapitel-pdf/270_brugger.pdf (visited on 12th Mar. 2016).
- (2009). 'Website history and the website as an object of study'. In: *New Media & Society* 11.1-2, pp. 115–132. URL: <https://doi.org/10.1177/1461444808099574>.
 - (2011). 'Web Archiving - Between Past, Present, and Future'. In: *The Handbook of Internet Studies*. Ed. by Mia Consalvo and Charles Ess. Oxford: Wiley-Blackwell, pp. 24–42.
 - (2012). 'When the Present Web is Later the Past: Web Historiography, Digital History, and Internet Studies.' In: *Historical Social Research* 37.4, pp. 102–117.
 - (2013). 'Web historiography and Internet Studies: Challenges and perspectives'. In: *New Media & Society* 15.5, pp. 752–764. URL: <https://doi.org/10.1177/1461444812462852>.
 - (2016). 'Digital Humanities in the 21st Century: Digital Material as a Driving Force'. In: *Digital Humanities Quarterly* 010.2.
 - (2018). *The Archived Web: Doing History in the Digital Age*. Cambridge, MA; London, England: MIT Press.
- Brügger, Niels and Ditte Laursen, eds. (2019). *The Historical Web and Digital Humanities: The Case of National Web Domains*. First. Abingdon, Oxon; New York, NY: Routledge.
- Brügger, Niels, Ditte Laursen and Janne Nielson (2017). 'Exploring the domain names of the Danish web'. In: *The Web as History: Using Web Archives to Understand the Past and the Present*. Ed. by Niels Brügger and Ralph Schroeder. London: UCL Press, pp. 62–80.
- Brügger, Niels and Ian Milligan, eds. (2019). *The SAGE Handbook of Web History*. London: SAGE.
- Brügger, Niels and Ralph Schroeder, eds. (2017). *The Web as History: Using Web Archives to Understand the Past and the Present*. London: UCL Press.

- Brunelle, Justin F. (2012). *Zombies in the Archives*. Web Science and Digital Libraries Research Group. URL: <https://ws-dl.blogspot.com/2012/10/2012-10-10-zombies-in-archives.html> (visited on 29th July 2019).
- Brunelle, Justin F. et al. (2015). 'Not all mementos are created equal: measuring the impact of missing resources'. In: *International Journal on Digital Libraries* 16.3, pp. 283–301. URL: <http://dx.doi.org/10.1007/s00799-015-0150-6>.
- Bruno, Isabelle, Emmanuel Didier and Tommaso Vitale (2014). 'Statactivism: Forms of action between disclosure and affirmation'. en. In: *Partecipazione e Conflitto* 7.2, pp. 198–220. ISSN: 2035-6609. DOI: 10.1285/i20356609v7i2p198. URL: <http://siba-esu.unisalento.it/index.php/paco/article/view/14150> (visited on 21st May 2019).
- Bucher, Taina (2012). 'Algorithmic Power and the Threat of Visibility on Facebook'. In: *New Media & Society* 14.7, pp. 1164–1180. URL: <https://doi.org/10.1177/1461444812440159>.
- Bueger, Christian (2014). 'Pathways to practice: praxiography and international politics'. In: *European Political Science Review* 6.3, pp. 383–406. DOI: <http://dx.doi.org/10.1017/S1755773913000167>.
- Bump, Philip (2018). 'The Joy Reid fight reinforces how critical the Internet Archive is to modern politics'. In: *Washington Post*. URL: <https://www.washingtonpost.com/news/politics/wp/2018/04/25/the-joy-reid-fight-reinforces-how-critical-the-internet-archive-is-to-modern-politics/> (visited on 13th Mar. 2019).
- Burgess, Jean and Axel Bruns (2012). 'Twitter Archives and the Challenges of "Big Social Data" for Media and Communication Research'. In: *M/C Journal* 15.5. URL: <http://www.journal.media-culture.org.au/index.php/mcjournal/article/view/561> (visited on 10th Aug. 2015).
- Butler, Chris (2018). *Addressing Recent Claims of "Manipulated" Blog Posts in the Wayback Machine*. Internet Archive Blogs. URL: <http://blog.archive.org/2018/04/24/addressing-recent-claims-of-manipulated-blog-posts-in-the-wayback-machine/> (visited on 14th Mar. 2019).
- Butler, Judith (1990). *Gender Trouble: Feminism and the Subversion of Identity*. First. New York and London: Routledge.
- (1996). 'Performativity's Social Magic'. In: *The Social and Political Body*. Ed. by Theodore R. Schatzki and Wolfgang Natter. New York: The Guilford Press.
- Byrne, Helena (2016). *2016 Rio Games Collection - How to Get Involved!* IIPC. URL: <https://netpreserveblog.wordpress.com/2016/06/27/2016-rio-games-collection-how-to-get-involved/> (visited on 9th Jan. 2017).
- Campbell, John (2018). *Removal of 26 Documents for Asylum Officer Training from the USCIS Website*. Access Assessment Report. Washington, D.C.: Sunlight Foundation's Web Integrity Project. URL: <http://sunlightfoundation.com/wp-content/uploads/2018/05/AAR-6-USCIS-Asylum-Training-Materials-180529.pdf> (visited on 12th July 2019).

- (2019). *USCIS Removed Asylum Training Documents from Website at Direction of Top Brass*. Sunlight Foundation. URL: <https://sunlightfoundation.com/2019/06/04/uscis-removed-asylum-training-documents-from-website-at-direction-of-top-brass/> (visited on 12th July 2019).
- Caswell, Michelle (2009). 'Irreparable Damage: Violence, Ownership, and Voice in an Indian Archive'. In: *Libri* 59.1, pp. 1–13. ISSN: 1865-8423. DOI: 10.1515/libr.2009.001. URL: <https://doi.org/10.1515/libr.2009.001> (visited on 12th May 2019).
- (2013). 'Not Just Between Us: A Riposte to Mark Greene'. In: *American Archivist* 76.2, pp. 605–808. URL: <https://americanarchivist.org/doi/pdf/10.17723/aarc.76.2.89324135v02r2q74> (visited on 18th May 2019).
- Charmaz, Kathy (2003). 'Grounded Theory: Objectivist and Constructivist methods'. In: *Strategies for Qualitative Inquiry*. Ed. by Norman K. Denzin and Yvonna S. Lincoln. Second. Thousand Oaks, Calif: Sage Publications, pp. 249–291.
- (2006). *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. London; Thousand Oaks, New Delhi: SAGE.
- Chenoweth, Erica and Jeremy Pressman (2017). 'This is what we learned by counting the women's marches'. In: *Washington Post*. ISSN: 0190-8286. URL: <https://www.washingtonpost.com/news/monkey-cage/wp/2017/02/07/this-is-what-we-learned-by-counting-the-womens-marches/> (visited on 21st May 2019).
- Cho, Alexander (2015a). 'Queer Reverb: Tumblr, Affect, Time'. In: *Networked Affect*. Ed. by Ken Hillis, Susanna Paasonen and Michael Petit. Cambridge, Massachusetts: MIT Press, pp. 43–57. URL: <https://ieeexplore.ieee.org/document/7087717>.
- (2015b). 'Sensuous Participation: Queer Youth of Color, Affect, and Social Media'. PhD. Austin, Texas: University of Texas at Austin. URL: <https://repositories.lib.utexas.edu/bitstream/handle/2152/31667/CHO-DISSERTATION-2015.pdf?sequence=1> (visited on 3rd Feb. 2019).
- Chun, Wendy Hui Kyong (2013). *Programmed Visions: Software and Memory*. Paperback. MIT Press.
- Clifford, James (1988). *The Predicament of Culture: Twentieth-Century Ethnography, Literature and Art*. Cambridge, MA: Harvard University Press.
- Clifford, James and George E. Marcus (1986). *Writing Culture: The Politics and Poetics of Ethnography*. Berkeley: University of California Press.
- 'Canada Watch Fall 2015: The Politics of Evidence' (2015). In: *Canada Watch*. Ed. by Colin Coates, p. 36. URL: https://yorkspace.library.yorku.ca/xmlui/bitstream/handle/10315/30190/CW_Fall2015_FINAL.pdf (visited on 8th May 2019).
- Cohen, Anthony P. (1985). *The Symbolic Construction of Community*. e-book (2001). London and New York: Routledge.
- Coleman, E. Gabriella (2013). *Coding Freedom: The Ethics and Aesthetics of Hacking*. Princeton, New Jersey: Princeton University Press.

- Coleman, Gabriella (2012). 'Anonymous'. In: *Depletion Design: A Glossary of Network Ecologies*. Ed. by Carolin Wiedemann and Soenke Zehle. Theory on Demand 8. Amsterdam: Institute of Network Cultures, pp. 11–16. URL: <https://networkcultures.org/blog/publication/no-8-depletion-design-a-glossary-of-network-ecologies-2/>.
- (2014). *Hacker, Hoaxer, Whistleblower, Spy: The Many Faces of Anonymous*. London and New York: Verso.
- Conroy, John (2008). *Alexa from A to Z: The Rise and Falter of a Web Pioneer*. CM-SWire.com. URL: <https://www.cmswire.com/cms/web-publishing/alexa-from-a-to-z-the-rise-and-falter-of-a-web-pioneer-002694.php> (visited on 25th July 2019).
- Cook, Terry (2001). 'Archival science and postmodernism: new formulations for old concepts'. In: *Archival Science* 1.1, pp. 3–24. URL: <http://dx.doi.org/10.1007/BF02435636> (visited on 12th Mar. 2016).
- (2007). 'Electronic Records, Paper Minds: The Revolution in Information Management and Archives in the Post-Custodial and Post-Modernist Era'. In: *Archives & Social Studies: A Journal of Interdisciplinary Research* 1.0, pp. 399–443. URL: http://archivo.cartagena.es/files/36-164-DOC_FICHER01/06-cook_electronic.pdf (visited on 11th Mar. 2016).
- (2013). 'Evidence, memory, identity, and community: four shifting archival paradigms'. In: *Archival Science* 13.2, pp. 95–120. URL: <http://dx.doi.org/10.1007/s10502-012-9180-7>.
- Cook, Terry and Joan M. Schwartz (2002). 'Archives, Records, and Power: From (Post-modern) Theory to (Archival) Performance'. In: *Archival Science* 2, pp. 171–185.
- Costa, Miguel and Mário J. Silva (2012). 'Evaluating Web Archive Search Systems'. In: *Web Information Systems Engineering - WISE 2012: 13th International Conference, Paphos, Cyprus, November 28-30, 2012. Proceedings*. Ed. by X. Sean Wang et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 440–454. URL: http://dx.doi.org/10.1007/978-3-642-35063-4_32.
- Currie, Morgan, Joan Donovan and Brittany Paris (2018). 'Preserving for a More Just Future: Tactics of Activist Data Archiving'. In: *Data Science Landscape: Towards Research Standards and Protocols*. Ed. by Usha Mujoo Munshi and Neeta Verma. Singapore: Springer Singapore, pp. 67–78. URL: https://doi.org/10.1007/978-981-10-7515-5_5.
- Currie, Morgan E. and Britt S. Paris (2018). 'Back-ups for the future: archival practices for data activism'. In: *Archives and Manuscripts* 46.2, pp. 124–142. ISSN: 0157-6895. URL: <https://doi.org/10.1080/01576895.2018.1468273> (visited on 30th Aug. 2018).
- Davidson, Joe (2017). 'Trump transition leader's goal is two-thirds cut in EPA employees'. In: *Washington Post*. URL: <https://www.washingtonpost.com/news/powerpost/wp/2017/01/30/trump-transition-leaders-goal-is-two-thirds-cut-in-epa-employees> (visited on 12th May 2019).

- Day Thomson, Sarah (2016). *Preserving Social Media*. Technology Watch Report 16-01. Great Britain: Digital Preservation Coalition, pp. 1–50. URL: <http://dx.doi.org/10.7207/twr16-01> (visited on 21st Mar. 2016).
- De Kosnik, Abigail (2016). *Rogue Archives: Digital Cultural Memory and Media Fandom*. Cambridge, Massachusetts; London, England: MIT Press.
- Dennis, Brady (2016a). ‘Scientists are frantically copying U.S. climate data, fearing it might vanish under Trump’. In: *Washington Post*. URL: <https://www.washingtonpost.com/news/energy-environment/wp/2016/12/13/scientists-are-frantically-copying-u-s-climate-data-fearing-it-might-vanish-under-trump/> (visited on 10th Apr. 2019).
- (2016b). ‘Trump: ‘I’m not a big believer in man-made climate change.’ In: *Washington Post*. URL: <https://www.washingtonpost.com/news/energy-environment/wp/2016/03/22/this-is-the-only-type-of-climate-change-donald-trump-believes-in/> (visited on 11th Apr. 2019).
- Derrida, Jacques (1998). *Archive Fever: A Freudian Impression*. Trans. by Eric Prenowitz. Paperback. Chicago; London: University of Chicago Press.
- Dewey, Caitlin (2014). ‘How Web archivists and other digital sleuths are unraveling the mystery of MH17’. In: *Washington Post*. ISSN: 0190-8286. URL: <https://www.washingtonpost.com/news/the-intersect/wp/2014/07/21/how-web-archivists-and-other-digital-sleuths-are-unraveling-the-mystery-of-mh17/> (visited on 10th July 2019).
- Dillon, Lindsey et al. (2017). ‘Environmental Data Justice and the Trump Administration: Reflections from the Environmental Data and Governance Initiative’. In: *Environmental Justice* 10.6, pp. 186–192. ISSN: 1939-4071. URL: <https://www.liebertpub.com/doi/10.1089/env.2017.0020> (visited on 23rd Apr. 2019).
- Dillon, Lindsey et al. (2019). ‘Situating Data in a Trumpian Era: The Environmental Data and Governance Initiative’. In: *Annals of the American Association of Geographers* 109.2, pp. 545–555. ISSN: 2469-4452. URL: <https://doi.org/10.1080/24694452.2018.1511410> (visited on 23rd Apr. 2019).
- DiMaggio, Paul et al. (2001). ‘Social Implications of the Internet’. In: *Annual Review of Sociology* 27.1, pp. 307–336. DOI: 10.1146/annurev.soc.27.1.307. URL: <https://doi.org/10.1146/annurev.soc.27.1.307>.
- Dougherty, Meghan (2007). ‘Archiving the Web: Collection, Documentation, Display, and Shifting Knowledge Production Paradigms’. PhD. Seattle, Washington, USA: University of Washington.
- (2014). ‘Property or Privacy? Reconfiguring Ethical Concerns Around Web Archival Research Methods’. In: *Selected Papers of Internet Research, 2014*. Association of Internet Researchers. URL: <http://spir.aoir.org/index.php/spir/article/view/735> (visited on 8th Feb. 2016).
- Dougherty, Meghan and Eric T. Meyer (2014). ‘Community, Tools, and Practices in Web Archiving: The State-of-the-Art in Relation to Social Science and Humanities

- Research Needs'. In: *Journal of the Association for Information Science and Technology* 65.11, pp. 2195–2209.
- Dougherty, Meghan et al. (2010). *Researcher Engagement with Web Archives State of the Art*. London: JISC.
- Dourish, Paul and Genevieve Bell (2011). *Divining a Digital Future: Mess and Mythology in Ubiquitous Computing*. Cambridge, MA: MIT Press.
- Downey, Gregory J. (2014). 'Making Media Work: Time, Space, Identity, and Labor in the Analysis of Information and Communication Infrastructures'. In: *Media Technologies: Essays on Communication, Materiality, and Society*. Ed. by Tarleton Gillespie, Pablo J. Boczkowski and Kirsten A. Foot. Cambridge, Massachusetts; London, England: MIT Press, pp. 141–165.
- Driscoll, Kevin and Shawn Walker (2014). 'Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data'. In: *International Journal of Communication* 8, pp. 1745–1764. ISSN: 1932-8036. URL: <https://ijoc.org/index.php/ijoc/article/view/2171/1159> (visited on 8th July 2019).
- Duguay, Stefanie (2018). *Why Tumblr's ban on adult content is bad for LGBTQ youth*. The Conversation. URL: <http://theconversation.com/why-tumblrs-ban-on-adult-content-is-bad-for-lgbtq-youth-108215> (visited on 4th Feb. 2019).
- Ecarma, Caleb (2018). *EXCLUSIVE: Joy Reid Claims Newly Discovered Homophobic Posts From Her Blog Were 'Fabricated'*. Mediaite. URL: <https://www.mediaite.com/online/exclusive-joy-reid-claims-newly-discovered-homophobic-posts-from-her-blog-were-fabricated/> (visited on 14th Mar. 2019).
- Edwards, Jane A. (2001). 'The Transcription of Discourse'. In: *The Handbook of Discourse Analysis*. Ed. by Deborah Schiffrin, Deborah Tannen and Heidi Hamilton. First. Malden, MA: Blackwell Publishing, pp. 321–348.
- Edwards, Paul N. (2013). *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Paperback. Cambridge, Massachusetts; London, England: MIT Press.
- Eichhorn, Kate (2008). 'Archival Genres: Gathering Texts and Reading Spaces'. In: *Invisible Culture* 12, pp. 1–10. URL: <http://ivc.lib.rochester.edu/archival-genres-gathering-texts-and-reading-spaces/>.
- (2013). *The Archival Turn in Feminism: Outrage in Order*. Philadelphia: Temple University Press.
- Electronic Publications Pilot Project Team and Electronic Collections Committee (1996). *Electronic Publications Pilot Project (EPPP): Final Report*. Tech. rep. Canada: Library and Archives Canada. URL: <http://www.nlc-bnc.ca/obj/p4/f2/e-report.pdf> (visited on 9th Jan. 2017).
- Eltgroth, Deborah R. (2009). 'Best Evidence and the Wayback Machine: Toward a Workable Authentication Standard for Archived Internet Evidence'. In: *Fordham Law Review* 78.1, pp. 181–215. URL: <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=4467&context=flr> (visited on 7th Feb. 2019).

- Environmental Data & Governance Initiative (2016a). *Seeding the Internet Archive's Web Crawler*. EDGI Data Preservation Guides. URL: <https://edgi-govdata-archiving.github.io/guides/seeding-internet-archive/> (visited on 15th May 2019).
- (2016b). *Understanding the Internet Archive Web Crawler*. EDGI Data Preservation Guides. URL: <https://edgi-govdata-archiving.github.io/guides/internet-archive-crawler/> (visited on 15th May 2019).
- Environmental Data and Governance Initiative (2017). *Everythingdotgov 1-Pager*. URL: <https://drive.google.com/file/d/0B2c5ivzmTMB0eUNEWDF3cnA5WVk/view> (visited on 28th May 2019).
- Ernst, Wolfgang (2014). 'Between the Archive and the Anarchivable'. In: *Mnemoscape* 1. URL: <https://www.mnemoscape.org/single-post/2014/09/04/Between-the-Archive-and-the-Anarchivable-by-Wolfgang-Ernst> (visited on 22nd Jan. 2017).
- Fairclough, Norman (2012). 'Critical discourse analysis'. In: *International Advances in Engineering and Technology (IAET)* 7, pp. 452–487. URL: <http://scholarism.net/FullText/2012071.pdf> (visited on 7th May 2019).
- Feuerstein, Adam (1999). *E-commerce loves Street: Critical Path plans encore*. San Francisco Business Times. URL: <https://www.bizjournals.com/sanfrancisco/stories/1999/05/24/newscolumn4.html> (visited on 29th July 2019).
- Findlay, Cassie (2011). *Where do old websites go to die? with Jason Scott of Archive Team - Podcast*. URL: <https://rkroundtable.org/2011/06/25/where-do-old-websites-go-to-die-with-jason-scott-of-archive-team-podcast/> (visited on 6th Dec. 2018).
- Fink, Marty and Quinn Miller (2014). 'Trans Media Moments: Tumblr, 2011-2013'. In: *Television & New Media* 15.7, pp. 611–626. URL: <http://journals.sagepub.com/doi/10.1177/1527476413505002> (visited on 2nd Feb. 2019).
- Finn, Megan (2018). *Documenting Aftermath: Information Infrastructures in the Wake of Disasters*. Cambridge, MA; London, England: MIT Press.
- Fleur, Nicholas St. (2017). 'Scientists, Feeling Under Siege, March Against Trump Policies'. In: *The New York Times*. ISSN: 0362-4331. URL: <https://www.nytimes.com/2017/04/22/science/march-for-science.html> (visited on 21st May 2019).
- Flinn, Andrew and Ben Alexander (2015). "Humanizing an inevitability political craft": Introduction to the special issue on archiving activism and activist archiving'. In: *Archival Science* 15.4, pp. 329–335. ISSN: 1573-7519. URL: <https://doi.org/10.1007/s10502-015-9260-6> (visited on 18th May 2019).
- Flinn, Andrew, Mary Stevens and Elizabeth Shepherd (2009). 'Whose memories, whose archives? Independent community archives, autonomy and the mainstream'. In: *Archival Science* 9.1, pp. 71–86. URL: <http://dx.doi.org/10.1007/s10502-009-9105-2>.

- Foo, Christopher (2013). *ArchiveTeam Warrior Infrastructure*. URL: https://www.archive.team.org/index.php?title=File:Archiveteam_warrior_infrastructure.png.
- Foot, Kirsten and Steven Schneider (2010). 'Object-Oriented Web Historiography'. In: *Web History*. Ed. by Niels Brügger. New York, NY: Peter Lang, pp. 61–79.
- Foot, Kirsten A. and Steven Schneider (2006). *Web Campaigning*. Boston: MIT Press.
- Foucault, Michel (1972). *Archaeology of Knowledge*. Psychology Press.
- Freelon, Deen (2018). 'Computational Research in the Post-API Age'. In: *Political Communication* 35.4, pp. 665–668. URL: <https://doi.org/10.1080/10584609.2018.1477506> (visited on 22nd Jan. 2019).
- Gambino, Lauren et al. (2017). 'Thousands protest against Trump travel ban in cities and airports nationwide'. In: *The Guardian*. ISSN: 0261-3077. URL: <https://www.theguardian.com/us-news/2017/jan/29/protest-trump-travel-ban-muslims-airports> (visited on 21st May 2019).
- Geertz, Clifford (1973). *The Interpretation of Cultures: Selected Essays*. New York: Basic Books.
- Gehl, Robert W. (2017). '(Critical) Reverse Engineering and Genealogy'. In: *Le foucauldien* 3.1. URL: <http://doi.org/10.16995/lefou.26>.
- Gerber, Rebecca (2019). *LibGuides: Copyright for Libraries: General Information*. American Library Association. URL: <https://libguides.ala.org/copyright> (visited on 28th July 2019).
- Gieryn, Thomas F. (1983). 'Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists'. In: *American Sociological Review* 48.6, pp. 781–795. ISSN: 0003-1224. URL: <https://www.jstor.org/stable/2095325> (visited on 12th May 2019).
- Gillespie, Tarleton (2013). *Tumblr, NSFW porn blogging, and the challenge of checkpoints*. Culture Digitally. URL: <http://culturedigitally.org/2013/07/tumblr-nsfw-porn-blogging-and-the-challenge-of-checkpoints/> (visited on 4th Feb. 2019).
- (2018). *Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media*. New Haven and London: Yale University Press.
- Gilliland-Swetland, Anne J. (2000). *Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment*. Technical Report 89. Washington, D.C.: Council on Library and Information Resources. URL: <https://www.clir.org/pubs/reports/pub89/pub89.pdf> (visited on 22nd Sept. 2016).
- Goel, Vinay (2016a). *Beta Wayback Machine - Now with Site Search!* Blog. URL: <https://blog.archive.org/2016/10/24/beta-wayback-machine-now-with-site-search/> (visited on 12th Jan. 2017).
- (2016b). *Crawl Priorities: December 2015 Milestone Report*. Tech. rep. Internet Archive.
- (2016c). *Defining Web pages, Web sites and Web captures*. Blog. URL: <http://blog.archive.org/2016/10/23/defining-web-pages-web-sites-and-web-captures/> (visited on 28th Dec. 2016).

- Goel, Vindu (2017). 'Verizon Completes \$4.48 Billion Purchase of Yahoo, Ending an Era'. In: *The New York Times*. URL: <https://www.nytimes.com/2017/06/13/technology/yahoo-verizon-marissa-mayer.html> (visited on 26th Feb. 2019).
- Gomes, Daniel and Miguel Costa (2014). 'The Importance of Web Archives for Humanities'. In: *International Journal of Humanities and Arts Computing* 8.1, pp. 106–123. URL: <http://dx.doi.org/10.3366/ijhac.2014.0122> (visited on 1st Feb. 2016).
- Gomes, Daniel, Sérgio Freitas and Mário J. Silva (2006). 'Design and Selection Criteria for a National Web Archive'. In: *Research and Advanced Technology for Digital Libraries*. Ed. by Julio Gonzalo et al. Vol. 4172. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 196–207. URL: http://dx.doi.org/10.1007/11863878_17.
- Gomes, Daniel, João Miranda and Miguel Costa (2011). 'A Survey on Web Archiving Initiatives'. In: *Research and Advanced Technology for Digital Libraries*. Ed. by Stefan Gradmann et al. Vol. 6966. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 408–420. URL: http://dx.doi.org/10.1007/978-3-642-24469-8_41.
- Gómez Cruz, Edgar and Elisenda Ardèvol (2013). 'Ethnography and the Field in Media(ted) Studies: A Practice Theory Approach'. In: *Westminster Papers in Communication and Culture* 9.3, pp. 27–46. URL: <http://www.westminsterpapers.org/articles/abstract/10.16997/wpcc.172/> (visited on 16th Oct. 2016).
- Gonzalez, John (2016). *20,000 Hard Drives on a Mission*. Internet Archive Blogs. URL: <https://blog.archive.org/2016/10/25/20000-hard-drives-on-a-mission/> (visited on 28th July 2019).
- Goodenough, Ward H. (1963). *Cooperation in Change: An Anthropological Approach to Community Development*. New York: Russell Sage Foundation.
- Gracy, Karen F. (2001). 'The Imperative to Preserve: Competing Definitions of Value in the World of Film Preservation'. PhD. Los Angeles: University of California.
- (2004). 'Documenting Communities of Practice: Making the Case for Archival Ethnography'. In: *Archival Science* 4.3, pp. 335–365. URL: <http://dx.doi.org/10.1007/s10502-005-2599-3>.
- Graham, Mark (2017). *Robots.txt meant for search engines don't work well for web archives*. Internet Archive Blogs. URL: <https://blog.archive.org/2017/04/17/robots-txt-meant-for-search-engines-dont-work-well-for-web-archives/> (visited on 29th July 2019).
- Gray, Gabriella and Scott Martin (2007). 'The UCLA Online Campaign Literature Archive: A case study'. In: *Proceedings of International Web Archiving Workshop '07*. Vancouver, BC, pp. 1–5. URL: http://iwaw.europarchive.org/07/IWAW2007_gray.pdf (visited on 11th Jan. 2017).
- Green, Heather (2002). *A Library as Big as the World*. News. URL: https://web.archive.org/web/20020601134105/http://www.businessweek.com/technology/content/feb2002/tc20020228_1080.htm (visited on 13th Jan. 2017).

- Greene, Mark (2013). 'A Critique of Social Justice as an Archival Imperative: What Is It We're Doing That's All That Important?' In: *The American Archivist* 76.2, pp. 302–334. ISSN: 0360-9081. DOI: 10.17723/aarc.76.2.147441214663kw43. URL: <https://www.americanarchivist.org/doi/abs/10.17723/aarc.76.2.147441214663kw43> (visited on 18th May 2019).
- Grotke, Abbie (2008). *2008 Member Profile Survey Results*. Tech. rep. IIPC. URL: <http://www.netpreserve.org/resources/2008-iipc-member-profile-survey-results> (visited on 3rd Oct. 2016).
- (2011a). *Ask the Recommending Officer: The Civil War Sesquicentennial Web Archive*. Blog. URL: <http://blogs.loc.gov/thesignal/2011/08/ask-the-recommending-officer-the-civil-war-sesquicentennial-web-archive/> (visited on 7th Jan. 2017).
- (2011b). 'Web Archiving at the Library of Congress'. In: *Computers in Libraries* 31.10. URL: <http://www.infoday.com/cilmag/dec11/Grotke.shtml> (visited on 20th Mar. 2016).
- Grotke, Abbie and Gina Jones (2010). 'DigiBoard: A Tool to Streamline Complex Web Archiving Activities at the Library of Congress'. In: *Proceedings of International Web Archiving Workshop IAWAW 2010*. Vienna, Austria, pp. 17–23. URL: <http://iawaw.europarchive.org/10/IAWAW2010.pdf> (visited on 11th Jan. 2017).
- Hale, Scott A., Grant Blank and Victoria D. Alexander (2017). 'Live versus archive: Comparing a web archive to a population of web pages'. In: *The Web as History: Using Web Archives to Understand the Past and the Present*. UCL Press, pp. 45–61. ISBN: 978-1-911307-42-6. URL: <http://www.jstor.org/stable/j.ctt1mtz55k.8>.
- Hale, Scott A. et al. (2014). 'Mapping the UK Webspace: Fifteen Years of British Universities on the Web'. In: *Proceedings of the 2014 ACM Conference on Web Science*. WebSci '14. New York, NY, USA: ACM, pp. 62–70. ISBN: 978-1-4503-2622-3. DOI: 10.1145/2615569.2615691. URL: <http://doi.acm.org/10.1145/2615569.2615691>.
- Halford, Susan, Cathy Pope and Leslie Carr (2010). 'A manifesto for Web Science?' In: *2010 Web Science Conference*. URL: <http://eprints.soton.ac.uk/271033/> (visited on 29th Nov. 2013).
- Hambridge, S. (1995). *Netiquette Guidelines*. RFC FYI:28. Network Working Group, Intel Corporation. URL: <https://www.rfc-editor.org/rfc/rfc1855.txt> (visited on 28th Feb. 2019).
- Hammersley, Martyn (2012). *What is Qualitative Research?* First. The 'What is?' Research Methods Series. London: Bloomsbury Publishing.
- Hammersley, Martyn and Paul Atkinson (2007). *Ethnography: Principles and Practice*. Third. London and New York: Routledge.
- Handwerker, W. Penn (2001). *Quick Ethnography*. Plymouth: AltaMira Press.
- Haraway, Donna (1988). 'Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective'. In: *Feminist Studies* 14.3, pp. 575–599.

- Hardiman, Rachel (2009). 'En mal d'archive: Postmodernist Theory and Recordkeeping'. In: *Journal of the Society of Archivists* 30.1, pp. 27–44. URL: <http://dx.doi.org/10.1080/00379810903264591>.
- Harris, Verne (2000). 'Law, Evidence and Electronic Records: Strategic Perspective from the Global Periphery'. In: *S. A. Archives Journal* 41, pp. 3–19.
- (2005). 'Archives, Politics, and Justice'. In: *Political Pressure and the Archival Record*. Ed. by Margaret Procter, Michael Cook and Caroline Williams. Chicago: Society of American Archivists, pp. 173–182.
- Hart, Matt (2015). 'Youth Intimacy on Tumblr: A Pilot Study'. In: *Young* 23.3, pp. 193–208. ISSN: 1103-3088. URL: <https://doi.org/10.1177/1103308815577878> (visited on 26th Feb. 2019).
- Herriott, Robert E. and William A. Firestone (1983). 'Multisite Qualitative Policy Research: Optimizing Description and Generalizability'. In: *Educational Researcher* 12.2, pp. 14–19. URL: <http://www.jstor.org/stable/1175416> (visited on 24th Sept. 2016).
- Hess, Joey (2015). *I am ArchiveTeam*. Blog. URL: https://joeyh.name/blog/entry/I_am_ArchiveTeam/ (visited on 7th Jan. 2017).
- Ho, Vivian (2018). 'Tumblr's adult content ban dismays some users: 'It was a safe space''. In: *The Guardian*. ISSN: 0261-3077. URL: <https://www.theguardian.com/technology/2018/dec/03/tumblr-adult-content-ban-lgbt-community-gender> (visited on 4th Feb. 2019).
- Hobsbawm, Eric (1972). 'Social Bandits: Reply'. In: *Comparative Studies in Society and History* 14.4, pp. 503–505. URL: <https://doi.org/10.1017/S0010417500006836> (visited on 28th Jan. 2019).
- Hobsbawm, Eric J. (1959). *Primitive Rebels: Studies in Archaic Forms of Social Movement in the 19th and 20th Centuries*. Manchester: University of Manchester Press.
- Hockx-Yu, Helen (2011). 'The Past Issue of the Web'. In: *Proceedings of the ACM WebSci'11*. Web Science 2011. Koblenz, Germany. URL: <http://www.websci11.org/fileadmin/websci/Papers/PastIssueWeb.pdf> (visited on 20th Mar. 2016).
- (2013). 'Archiving Social Media in the Context of Non-print Legal Deposit'. In: *IFLA 2013 Proceedings*. Lyon: IFLA, pp. 1–10. URL: <http://library.ifla.org/999/1/107-hockxyu-en.pdf> (visited on 20th Mar. 2016).
- (2014a). 'Access and Scholarly Use of Web Archives'. In: *Alexandria: The Journal of National and International Library and Information Issues* 25.1-2, pp. 113–127. URL: <https://doi.org/10.7227/ALX.0023> (visited on 8th Feb. 2016).
- (2014b). *OpenWayBack: General Overview*. GitHub. URL: <https://github.com/iipc/openwayback/wiki/General-overview> (visited on 27th Jan. 2017).
- Hogan, Lynn (1995). 'Creating a Web Page Using AOL Hometown'. In: *Practical Computing*. Pearson Education. URL: https://web.archive.org/web/20170130091141/http://wps.prenhall.com/bp_hogan_webchapters_1/23/6004/1537044.cw/content/index.html (visited on 30th Jan. 2017).

- Hu, Jim (2002). *AOL home page glitches irk users*. Magazine. URL: <https://www.cnet.com/news/aol-home-page-glitches-irk-users/> (visited on 10th Jan. 2017).
- Huc-Hepher, Saskia (2015). 'Big Web data, small focus: An ethnosemiotic approach to culturally themed selective Web archiving'. In: *Big Data & Society* 2.2, pp. 1–15. DOI: 10.1177/2053951715595823. URL: <https://doi.org/10.1177/2053951715595823>.
- Hughes, Everett C. (1971). *The Sociological Eye*. Chicago and New York: Aldine and Atherton.
- IIPC (2004). *International Internet Preservation Consortium*. Institutional. URL: <https://web.archive.org/web/20120501092823/http://netpreserve.org/press/pr20040505.php> (visited on 15th Apr. 2016).
- Internet Memory Foundation (2010). *Web Archiving in Europe*. Tech. rep. Internet Memory Foundation. URL: http://internetmemory.org/images/uploads/Web_Archiving_Survey.pdf (visited on 13th Mar. 2016).
- ISO (2009). *ISO 28500:2009: Information and documentation – WARC file format*. URL: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44717 (visited on 11th Mar. 2016).
- Ivanov, Asen Ognyanov (2017). 'Practice theory: a new approach for archival and recordkeeping research'. In: *Records Management Journal* 27.2, pp. 104–124. URL: <https://doi.org/10.1108/RMJ-10-2016-0038>.
- Ivey, Paul Eli (1999). *Prayers in stone: Christian Science architecture in the United States*. Urbana and Chicago: University of Illinois Press.
- Jackson, Steven J. (2014). 'Rethinking Repair'. In: *Media Technologies: Essays on Communication, Materiality, and Society*. Ed. by Tarleton Gillespie, Pablo J. Boczkowski and Kirsten A. Foot. Cambridge, Massachusetts; London, England: MIT Press, pp. 221–239.
- Jacobs, James and Jefferson Bailey (2017). *The End of Term Web Archive: Collecting & Preserving the .gov Information Sphere*. San Jose State University. URL: <https://scholarworks.sjsu.edu/slasc/15> (visited on 29th Nov. 2018).
- Jasanoff, Sheila (2004). *States of Knowledge: The Co-Production of Science and Social Order*. London: Routledge.
- Jasanoff, Sheila and Sang-Hyun Kim (2009). 'Containing the Atom: Sociotechnical Imaginaries and Nuclear Power in the United States and South Korea'. In: *Minerva* 47.2, p. 119. URL: <https://doi.org/10.1007/s11024-009-9124-4>.
- Jeff (2002). *exclusions from the Wayback Machine*. Internet Archive - Web Forum. URL: <http://web.archive.org/web/20021001125048/http://www.archive.org:80/iathreads/post-view.php?id=778> (visited on 29th July 2019).
- Jewett, Tom and Rob Kling (1991). 'The Dynamics of Computerization in a Social Science Research Team: A Case Study of Infrastructure, Strategies, and Skills'. In: *Social Science Computer Review* 9.2, pp. 246–275. ISSN: 0894-4393. DOI: 10.1177/089443939100900205. URL: <https://doi.org/10.1177/089443939100900205> (visited on 26th July 2019).

- Johare, Rusnah and Mohamad Noorman Masrek (2011). 'Malaysian archival heritage at risk?: A survey of archivists' knowledge and skills in managing electronic records'. In: *Library Review* 60.8, pp. 685–711. URL: <http://dx.doi.org/10.1108/00242531111166719>.
- Kahle, Brewster (1997). 'Preserving the Internet'. In: *Scientific American* 276.3, pp. 82–83.
- (2007). *Internet Archive officially a library*. URL: <https://archive.org/post/121377/internet-archive-officially-a-library> (visited on 3rd Feb. 2017).
 - (2013). *Blacked Out Government Websites Available Through Wayback Machine*. In: Internet Archive Blogs. URL: <https://blog.archive.org/2013/10/02/governmentblackout/> (visited on 9th July 2019).
- Kahle, Brewster and Ana Parejo Vardillo (2015). 'The Internet Archive: An Interview with Brewster Kahle'. In: *19: Interdisciplinary Studies in the Long Nineteenth Century* 2015.21. ISSN: 1755-1560. DOI: 10.16995/ntn.760. URL: <http://doi.org/10.16995/ntn.760> (visited on 7th Apr. 2019).
- Kahle, Brewster et al. (1992). 'Wide Area Information Servers: An Executive Information System for Unstructured Files'. In: *Internet Research*. DOI: 10.1108/eb047255. URL: <https://www.emerald.com/insight/content/doi/10.1108/eb047255/full/html> (visited on 27th July 2019).
- Karpf, David (2012). 'Social Science Research Methods in Internet Time'. In: *Information, Communication & Society* 15.5, pp. 639–661. URL: <https://doi.org/10.1080/1369118X.2012.665468>.
- Kelleher, Christian (2017). 'Archives Without Archives: (Re)Locating and (Re)Defining the Archive Through Post-Custodial Praxis'. In: *Journal of Critical Library and Information Studies* 1.2. DOI: 10.24242/jclis.v1i2.29. URL: <https://doi.org/10.24242/jclis.v1i2.29> (visited on 31st July 2019).
- Kelly, Ryan (2009). *Twitter Study*. White Paper. Pear Analytics. URL: <http://pearanalytics.com/wp-content/uploads/2009/08/Twitter-Study-August-2009.pdf> (visited on 8th Apr. 2016).
- Kelty, Christopher M. (2008). *Two Bits: The Cultural Significance of Free Software*. Durham, NC: Duke University Press.
- Khoo, Michael, Lily Rozaklis and Catherine Hall (2012). 'A survey of the use of ethnographic methods in the study of libraries and library users'. In: *Library & Information Science Research* 34.2, pp. 82–91. URL: <https://doi.org/10.1016/j.lisr.2011.07.010>.
- Kim, Heejung and Hyewon Lee (2009). 'Digital-age trends and perspectives in Korean university archives'. In: *The Electronic Library* 27.3, pp. 426–440. URL: <http://dx.doi.org/10.1108/02640470910966871>.
- Kimpton, Michele and Jeff Ubois (2006). 'Year-by-Year: From an Archive of the Internet to an Archive on the Internet'. In: *Web Archiving*. Ed. by Julien Masanès. First. Berlin, Heidelberg: Springer, pp. 201–212.

- Kitchin, Rob (2016). 'Thinking critically about and researching algorithms'. In: *Information, Communication & Society* 20.1, pp. 14–29. URL: <http://dx.doi.org/10.1080/1369118X.2016.1154087>.
- Kitschelt, Herbert P. (1986). 'Political Opportunity Structures and Political Protest: Anti-Nuclear Movements in Four Democracies'. In: *British Journal of Political Science* 16.1, pp. 57–85. ISSN: 0007-1234. URL: <https://www.jstor.org/stable/193981> (visited on 21st May 2019).
- Klein, Martin et al. (2014). 'Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot'. In: *PLOS ONE* 9.12, pp. 1–39. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0115253. URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115253> (visited on 8th July 2019).
- Knoblauch, Hubert (2005). 'Focused Ethnography'. In: *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 6.3. URL: <http://www.qualitative-research.net/index.php/fqs/article/view/20>.
- Knorr Cetina, Karin (1999). *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, MA; London, England: Harvard University Press.
- Knuth, Rebecca (2003). *Libricide: The Regime-sponsored Destruction of Books and Libraries in the Twentieth Century*. Westport, Connecticut; London: Praeger.
- Koebler, Jason (2016). 'Researchers Are Preparing for Trump to Delete Government Science From the Web'. en-US. In: *Vice*. URL: https://www.vice.com/en_us/article/jpgnvx/researchers-are-preparing-for-trump-to-delete-government-science-from-the-web (visited on 21st May 2019).
- Koebler, Jason and Samantha Cole (2018). *Apple Sucked Tumblr Into Its Walled Garden, Where Sex Is Bad*. Motherboard. URL: https://motherboard.vice.com/en_us/article/a3mjxg/apple-tumblr-porn-nsfw-adult-content-banned (visited on 4th Feb. 2019).
- Koehler, Wallace (2004). 'A longitudinal study of Web pages continued: a consideration of document persistence'. In: *Information Research* 9.2. URL: <http://www.informationr.net/ir/9-2/paper174.html> (visited on 26th Feb. 2016).
- Koster, Martijn (1993). *Guidelines for Robot Writers*. URL: <http://www.robotstxt.org/guidelines.html> (visited on 14th Apr. 2016).
- Kozinets, Robert V. (2010). *Netnography: Doing Ethnographic Research Online*. London: SAGE.
- Kozłowska, Hanna (2018). *Tumblr is banning porn and other adult content*. Quartz. URL: <https://qz.com/1482821/tumblr-is-banning-porn-and-other-adult-content/> (visited on 4th Feb. 2019).
- Kuny, Terry (1997). 'A Digital Dark Ages? Challenges in the Preservation of Electronic Information'. In: *Proceedings of the 63rd International Federation of Library Associations and Institutions*. Copenhagen, Denmark.
- Kupferman, Steve (2016). 'Q&A: Michelle Murphy, the U of T professor who's racing to preserve climate-change data before Donald Trump takes office'. In: *Toronto Life*. URL: <https://torontolife.com/city/toronto-politics/qa-michelle->

- murphy - u - t - professor - whos - racing - preserve - climate - change - data - donald-trump-takes-office/ (visited on 24th Apr. 2019).
- Ladd, Kelly (2009). 'Textuality, Performativity and Archive: Examining the Virtual Body in Socially Networked Space'. Masters of Arts. Toronto, Canada: University of Toronto. URL: https://tspace.library.utoronto.ca/bitstream/1807/18093/3/Ladd_Kelly_200911_MA_thesis.pdf (visited on 15th Mar. 2015).
- Lamdan, Sarah (2018). 'Lessons from DataRescue: The Limits of Grassroots Climate Change Data Preservation and the Need for Federal Records Law Reform'. In: *University of Pennsylvania Law Review Online* 166.231, pp. 231–248. ISSN: 1942-8537. URL: <https://www.pennlawreview.com/online/166-U-Pa-L-Rev-Online-231.pdf>.
- LANIGAN, CLARE (2015). *Archiving Tweets: Reckoning with Twitter's Policy*. News. URL: <http://newslab.insight-centre.org/tweetarchivingchallenges/> (visited on 14th Apr. 2016).
- Latour, Bruno (1987). *Science in Action: How to Follow Scientists and Engineers through Society*. Cambridge, MA: Harvard University Press.
- (2005). *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.
- Law, John (1990). 'Introduction: Monsters, Machines and Sociotechnical Relations'. In: *The Sociological Review* 38.1, pp. 1–23. ISSN: 0038-0261. DOI: 10.1111/j.1467-954X.1990.tb03346.x. URL: <https://doi.org/10.1111/j.1467-954X.1990.tb03346.x> (visited on 16th July 2019).
- Law, John and John Hassard, eds. (1999). *Actor Network Theory and After*. Oxford: Blackwell and The Sociological Review.
- Lee, Martha F. (1995). *Earth First!: Environmental Apocalypse*. Syracuse, New York: Syracuse University Press.
- Leetaru, Kalev (2015). *Why It's So Important To Understand What's In Our Web Archives*. News. URL: <http://onforb.es/1VDPHPH> (visited on 28th Feb. 2016).
- (2016). *The Internet Archive Turns 20: A Behind The Scenes Look At Archiving The Web*. News. URL: <http://www.forbes.com/sites/kalevleetaru/2016/01/18/the-internet-archive-turns-20-a-behind-the-scenes-look-at-archiving-the-web/#747db6257800> (visited on 30th Jan. 2017).
- Levy, Steven (2010). *Hackers: Heroes of the Computer Revolution*. Sebastopol, CA: O'Reilly.
- Liamputtong, Pranee (2007). *Researching the Vulnerable: A Guide to Sensitive Research Methods*. London: SAGE.
- Library of Congress (2013). *Update on the Twitter Archive at the Library of Congress*. White Paper. Washington, D.C.: Library of Congress, pp. 1–5. URL: http://www.loc.gov/today/pr/2013/files/twitter_report_2013jan.pdf (visited on 15th Apr. 2016).
- Lievrouw, Leah A. (2014). 'Materiality and Media in Communication and Technology Studies: An Unfinished Project'. In: *Media Technologies: Essays on Communication*,

- Materiality, and Society*. Ed. by Tarleton Gillespie, Pablo J. Boczkowski and Kirsten A. Foot. Cambridge, Massachusetts; London, England: MIT Press, pp. 21–51.
- Lievrouw, Leah A. and Sonia Livingstone (2006). 'Introduction to the Updated Student Edition'. In: *Handbook of New Media: Social Shaping and Social Consequences*. Ed. by Leah A. Lievrouw and Sonia Livingstone. Fully revised student edition. London, UK: Sage Publications, pp. 1–14. ISBN: 978-1-4129-1873-2. URL: <http://eprints.lse.ac.uk/id/eprint/21502> (visited on 18th July 2019).
- Livingston, Jessica (2007). *Founders at Work: Stories of Startups' Early Days*. United States of America: Apress.
- Locker, Melissa (2018). *Tumblr's adult flagging tech is not working very well unless you think Sesame Street is NSFW*. Fast Company. URL: <https://www.fastcompany.com/90278880/tumblrs-adult-flagging-tech-is-not-working-very-well-unless-you-think-sesame-street-is-nsfw> (visited on 24th Jan. 2019).
- Lor, Peter and Johannes J. Britz (2004). 'A moral perspective on South-North web archiving'. In: *Journal of Information Science* 30.6, pp. 540–549. URL: <https://doi.org/10.1177/0165551504047925> (visited on 2nd Feb. 2016).
- (2012). 'An ethical perspective on political-economic issues in the long-term preservation of digital heritage'. In: *Journal of the American Society for Information Science and Technology* 63.11, pp. 2153–2164. URL: <http://dx.doi.org/10.1002/asi.22725> (visited on 20th Mar. 2016).
- Lunden, Ingrid (2013). *Yahoo: Expect Ads On Tumblr To Ramp Up Significantly In 2014*. Tech Crunch. URL: <https://techcrunch.com/2013/05/20/yahoo-expect-ads-on-tumblr-to-ramp-up-significantly-in-2014/> (visited on 14th Feb. 2019).
- Lyman, Peter (2002). 'Archiving the World Wide Web'. In: *Building a National Strategy for Preservation: Issues in Digital Media Archiving*. Council on Library, Information Resources and the Library of Congress, pp. 38–51. URL: <http://www.clir.org/pubs/reports/pub106/web.html> (visited on 13th Aug. 2015).
- Lyotard, Jean-Francois (1984). *The Postmodern Condition: A Report on Knowledge*. Trans. by Geoff Bennington and Brian Massumi. English Translation. Theory and History of Literature 10. Manchester: Manchester University Press.
- Madden, Raymond (2010). *Being Ethnographic*. Thousand Oaks, CA: SAGE.
- Maemura, Emily et al. (2018). 'If these crawls could talk: Studying and documenting web archives provenance'. In: *Journal of the Association for Information Science and Technology* 69.10, pp. 1223–1233. ISSN: 2330-1643. URL: <https://doi.org/10.1002/asi.24048> (visited on 12th Mar. 2019).
- Manovich, Lev (2012). 'Trending: the promises and the challenges of big social data'. In: *Debates in the Digital Humanities*. Ed. by M. K. Gold. Minneapolis, Minnesota: The University of Minnesota Press. URL: <http://manovich.net/content/04-projects/067-trending-the-promises-and-the-challenges-of-big-social-data/64-article-2011.pdf> (visited on 8th Apr. 2016).

- Marche, Stephen (2015). 'The Closing of the Canadian Mind'. In: *The New York Times*. URL: <https://www.nytimes.com/2015/08/16/opinion/sunday/the-closing-of-the-canadian-mind.html> (visited on 12th May 2019).
- Marcus, George E. (1995). 'Ethnography In/Of the World System: The Emergence of Multi-Sited Ethnography'. In: *Annual Review of Anthropology* 24, pp. 95–117.
- Marres, Noortje and Esther Weltevrede (2013). 'Scraping the Social?' In: *Journal of Cultural Economy* 6.3, pp. 313–335. URL: <http://dx.doi.org/10.1080/17530350.2013.772070> (visited on 9th Apr. 2016).
- Masanès, Julien, ed. (2006). *Web Archiving*. First. Springer-Verlag Berlin Heidelberg.
- Mason, Ingrid (2007). 'Virtual preservation: How has digital culture influenced our ideas about permanence? changing practice in a national legal deposit library.' In: *Library Trends* 56.1, pp. 198–215.
- Mason, Jennifer (2011). 'Facet Methodology: The Case for an Inventive Research Orientation'. In: *Methodological Innovations Online* 6.3, pp. 75–92. ISSN: 1748-0612. DOI: 10.4256/mio.2011.008. URL: <https://journals.sagepub.com/doi/abs/10.4256/mio.2011.008> (visited on 15th Nov. 2018).
- Mayer, Marissa (2013). *Tumblr + Yahoo! = !!* Blog. URL: <https://yahoo.tumblr.com/post/50902111638/tumblr-yahoo> (visited on 26th Feb. 2019).
- McCarthy, Kieren (2018). *Archive.org's Wayback Machine is legit legal evidence, US appeals court judges rule*. The Register. URL: https://www.theregister.co.uk/2018/09/04/wayback_machine_legit/ (visited on 30th Jan. 2019).
- Mckemmish, Sue and Anne Gilliland (2013). 'Archival and recordkeeping research: Past, present and future'. In: *Research Methods: Information, Systems and Contexts*. Ed. by K. Williamson and G. Johanson. Prahran, Victoria: Tilde Publishing, pp. 79–112.
- McKemmish, Sue, Franklyn Herbert Upward and Barbara Reed (2010). 'Records Continuum Model'. In: *Encyclopedia of Library and Information Sciences*. Ed. by Marcia J. Bates and Mary Niles Maack. Third. New York: Taylor & Francis, pp. 4447–4459.
- Merriam, Sharan B. (2009). *Qualitative Research: A guide to design and implementation*. Third. San Francisco, California: Jossey-Bass.
- Merriam, Sharan B. and Elizabeth J. Tisdell (2016). *Qualitative Research: A guide to design and implementation*. Fourth. San Francisco, California: Jossey-Bass.
- Meyer, Eric T., Arthur Thomas and Ralph Schroeder (2011). *Web Archives: The Future(s)*. Tech. rep. Oxford: Oxford Internet Institute, University of Oxford. URL: <http://ssrn.com/paper=1830025> (visited on 17th Feb. 2016).
- Meyer, Eric T. et al. (2017). 'Analysing the UK web domain and exploring 15 years of UK universities on the web'. In: *The Web as History: Using Web Archives to Understand the Past and Present*. Ed. by Neils Brügger and Ralph Schroeder. London: UCL Press.
- Mieszkowski, Katharine (2001). *Dumpster diving on the Web*. Magazine. URL: <https://web.archive.org/web/20170113090339/http://www.salon.com/2001/11/02/wayback/> (visited on 13th Jan. 2017).

- Milan, Stefania (2016). 'Data Activism as the New Frontier of Media Activism'. en. In: *Media Activism in the Digital Age: Charting an Evolving Field of Research*. Ed. by Goubin Yang and Viktor Pickard. Rochester, NY: Routledge. URL: <https://papers.ssrn.com/abstract=2882030> (visited on 21st May 2019).
- Miles, Matthew B., A. Michael Huberman and Johnny Saldana (2014). *Qualitative Data Analysis: A Methods Sourcebook*. Third. Thousand Oaks, CA: SAGE.
- Millar, Laura (2017). 'On the crest of a wave: transforming the archival future'. In: *Archives and Manuscripts* 45.2, pp. 59–76. ISSN: 0157-6895. DOI: 10.1080/01576895.2017.1328696. URL: <https://doi.org/10.1080/01576895.2017.1328696> (visited on 30th July 2019).
- Miller, Daniel and Don Slater (2000). *The Internet: an ethnographic approach*. Oxford: Berg.
- Milligan, Ian (2016). 'Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives'. In: *International Journal of Humanities and Arts Computing* 10.1, pp. 78–94. URL: <http://dx.doi.org/10.3366/ijhac.2016.0161>.
- (2017). 'Welcome to the web: The online community of GeoCities during the early years of the World Wide Web'. In: *The Web as History: Using Web Archives to Understand the Past and Present*. Ed. by Niels Brügger and Ralph Schroeder. London: UCL Press, pp. 137–158.
- (2019). *History in the Age of Abundance*. Montreal & Kingston; London; Chicago: McGill-Queen's University Press.
- Milligan, Ian, Nick Ruest and Jimmy Lin (2016). 'Content Selection and Curation for Web Archiving: The Gatekeepers vs. the Masses'. In: *JCDL '16, June 19 - 23, 2016, Newark, NJ, USA*. Newark, NJ: ACM. DOI: <http://dx.doi.org/10.1145/2910896.2910913>.
- Mohr, Gordon et al. (2004). 'An Introduction to Heritrix: An open source archival quality web crawler'. In: *Proceedings of the 4th International Web Archiving Workshop*. Bath, UK. URL: <https://webarchive.jira.com/wiki/download/attachments/5441/Mohr-et-al-2004.pdf> (visited on 3rd Mar. 2016).
- Mol, Annemarie (2002). *The Body Multiple: Ontology in Medical Practice*. Durham and London: Duke University Press.
- Morgan, Gareth (1997). *Images of Organisation*. Second. Thousand Oaks, CA: SAGE Publications.
- Morita, Atsuro (2014). 'The Ethnographic Machine: Experimenting with Context and Comparison in Strathernian Ethnography'. In: *Science, Technology, & Human Values* 39.2, pp. 214–235.
- Mortillaro, Nicole (2016). *U of T heads 'guerrilla archiving event' to preserve climate data ahead of Trump presidency*. CBC. URL: <https://www.cbc.ca/news/technology/university-toronto-guerrilla-archiving-event-trump-climate-change-1.3896167> (visited on 21st May 2019).

- Murdoch, Frances Russell (2010). 'Harper's Latest Step in Building "Tea Party North"'. In: *The Tyee*. URL: <http://thetyee.ca/Opinion/2010/08/12/TeaPartyNorth/> (visited on 12th May 2019).
- Murphy, Michelle (2016). *Guerrilla Archiving Event: Saving Environmental Data from Trump*. Technoscience Research Unit. URL: <https://technoscienceunit.org/2016/12/04/guerrilla-archiving-event-saving-environmental-data-from-trump/> (visited on 14th May 2019).
- Nadai, Eva and Christoph Maeder (2005). 'Fuzzy Fields. Multi-Sited Ethnography in Sociological Research'. In: *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 6.3. URL: <http://www.qualitative-research.net/index.php/fqs/article/view/22/47> (visited on 23rd Sept. 2016).
- Nagel, Thomas (1986). *The View from Nowhere*. Oxford: Oxford University Press.
- National Archives (2011). *Web Archiving Guidance*. Tech. rep. London: The National Archives. URL: <https://nationalarchives.gov.uk/documents/information-management/web-archiving-guidance.pdf> (visited on 20th Mar. 2016).
- (2014). *UK Government Web Archive captures official tweets and videos*. Institutional. URL: <http://www.nationalarchives.gov.uk/news/929.htm> (visited on 20th Mar. 2016).
- National Archives and Records Administration (2008). *National Archives and Records Administration Web Harvest Background Information*. URL: <https://www.archives.gov/files/records-mgmt/pdf/nwm13-2008-brief.pdf> (visited on 22nd May 2019).
- National Digital Stewardship Alliance (2012). *Web Archiving Survey Report*. Tech. rep. NDSA Content Working Group. URL: http://www.digitalpreservation.gov/documents/ndsa_web_archiving_survey_report_2012.pdf (visited on 3rd Oct. 2016).
- National Library of Australia (2005). *Online Australian Publications: Selection Guidelines for Archiving and Preservation by the National Library of Australia*. Institutional. URL: <http://pandora.nla.gov.au/selectionguidelines.html> (visited on 20th Mar. 2016).
- NDIIPP (2012). *Science@Risk: Toward a National Strategy for Preserving Online Science*. Tech. rep. Washington, D.C.: National Digital Information Infrastructure and Preservation Program, pp. 1–35. URL: <http://1.usa.gov/1U8ccwA> (visited on 20th Mar. 2016).
- Nicolini, Davide (2012). *Practice Theory, Work, and Organization: An Introduction*. First. Oxford: Oxford University Press.
- Nisbett, Richard E. and Timothy DeCamp Wilson (1977). 'Telling More Than We Can Know: Verbal Reports on Mental Processes'. In: *Psychological Review* 84.3, pp. 231–259.
- Niu, Jinfang (2012). 'An Overview of Web Archiving'. In: *D-Lib Magazine* 18.3/4. URL: <http://www.dlib.org/dlib/march12/niu/03niu1.html> (visited on 24th Feb. 2016).

- Nocera, José L. Abdelnour (2002). 'Ethnography and Hermeneutics in Cybercultural Research Accessing IRC Virtual Communities'. In: *Journal of Computer-Mediated Communication* 7.2, pp. 0–0. URL: <http://dx.doi.org/10.1111/j.1083-6101.2002.tb00146.x>.
- Nollinger, Mark (1995). 'America, Online!' In: *Wired*. ISSN: 1059-1028. URL: <https://www.wired.com/1995/09/aol-2/> (visited on 27th July 2019).
- Noordegraaf, Julia (2011). 'Remembering the Past in the Dynarchive: The State of Knowledge in Digital Archives'. In: *Media in Transition 7: Unstable Platforms: The Promise and Peril of Transition, May 13-15, 2011*. Boston, MA: MIT. URL: <http://web.mit.edu/comm-forum/mit7/papers/Noordegraaf.pdf> (visited on 22nd Jan. 2017).
- Nwala, Alexander C., Michele C. Weigle and Michael L. Nelson (2018). 'Scraping SERPs for Archival Seeds: It Matters When You Start'. In: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. JCDL '18. New York, NY, USA: ACM, pp. 263–272. ISBN: 978-1-4503-5178-2. DOI: 10.1145/3197026.3197056. URL: <http://doi.acm.org/10.1145/3197026.3197056>.
- Ogden, Jessica, Susan Halford and Les Carr (2017). 'Observing Web Archives: The Case for an Ethnographic Study of Web Archiving'. In: *Proceedings of WebSci'17*. Web Science. Troy, NY USA: ACM. DOI: <https://doi.org/10.1145/3091478.3091506>.
- Oguz, Fatih and Wallace Koehler (2015). 'URL decay at year 20: A research note'. In: *Journal of the Association for Information Science and Technology* 67.2, pp. 477–479. URL: <http://dx.doi.org/10.1002/asi.23561> (visited on 27th Feb. 2016).
- Oldenburg, Ray (1989). *The Great Good Place: Cafes, Coffee Shaps, Bookstores, Bars, Hair Salons, and Other Hangouts at the Heart of a Community*. New York: Marlow & Company.
- O'Reilly, Karen (2012). *Ethnographic Methods*. Second. London: Routledge.
- (2015). 'Ethnography: Telling Practice Stories'. In: *Emerging Trends in the Social and Behavioral Sciences*. John Wiley & Sons, Inc. URL: <http://dx.doi.org/10.1002/9781118900772.etrds0120>.
- Ormerod, T. C. et al. (2005). 'Mixing Research Methods in HCI: Ethnography Meets Experimentation in Image Browser Design'. In: *Engineering Human Computer Interaction and Interactive Systems: Joint Working Conferences EHCI-DSVIS 2004, Hamburg, Germany, July 11-13, 2004, Revised Selected Papers*. Ed. by Rémi Bastide, Philippe Palanque and Jörg Roth. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 112–128. URL: http://dx.doi.org/10.1007/11431879_7.
- Ortner, Sherry B. (1984). 'Theory in Anthropology since the Sixties'. In: *Comparative Studies in Society and History* 26.1, pp. 126–166. URL: <http://www.jstor.org/stable/178524> (visited on 16th Oct. 2016).
- (2006). *Anthropology and Social Theory: Culture, Power, and the Acting Subject*. Durham and London: Duke University Press.

- Owens, Trevor (2012). *ArtBase and the Conservation and Exhibition of Born Digital Art: An Interview with Ben Fino-Radin*. In collab. with Ben Fino-Radin. URL: <http://blogs.loc.gov/thesignal/2012/05/artbase-and-the-conservation-and-exhibition-of-born-digital-art-an-interview-with-ben-fino-radin/> (visited on 7th Jan. 2017).
- (2014). *What Do you Mean by Archive? Genres of Usage for Digital Preservers* | *The Signal*. URL: <http://blogs.loc.gov/thesignal/2014/02/what-do-you-mean-by-archive-genres-of-usage-for-digital-preservers/> (visited on 20th June 2020).
- Pelto, Pertti (2013). *Applied Ethnography: Guidelines for Field Research*. Walnut Creek, California: Left Coast Press Inc.
- Penn Program in Environmental Humanities (n.d.). *DataRescue Events*. PPEH Lab. URL: <http://web.archive.org/web/20171211214439/http://www.ppehlab.org/datarescue-events/> (visited on 4th June 2019).
- Pennock, Maureen (2013). *Web-Archiving*. Tech. rep. Technology Watch Report 13:01. Great Britain: Digital Preservation Coalition, pp. 1–50. URL: <http://dx.doi.org/10.7207/twr13-01> (visited on 27th Feb. 2016).
- Peters, John Durham (2015). *The Marvelous Clouds: Toward a Philosophy of Elemental Media*. Chicago: University of Chicago Press.
- Phillips, Margaret E. (2005). ‘What Should We Preserve? The Question for Heritage Libraries in a Digital World’. In: *Library Trends* 54.1, pp. 57–71. URL: http://muse.jhu.edu/journals/library_trends/v054/54.1phillips.html (visited on 9th Feb. 2016).
- Phillips, Mark E. and Kristy K. Phillips (2019). ‘End of Term 2016 Presidential Web Archive’. en. In: *Against the Grain* 29.6, Article 10, pp. 27–30. ISSN: 2380-176X. URL: <https://doi.org/10.7771/2380-176X.7874> (visited on 3rd Feb. 2019).
- Pinch, Steven and N Henry (1999). ‘Discursive Aspects of Technological Innovation: The Case of the British Motor-Sport Industry’. In: *Environment and Planning A* 31.4, pp. 665–682. URL: <https://doi.org/10.1068/a310665>.
- Pollner, Melvin and R. M. Emerson (1983). ‘The Dynamics of Inclusion and Distance in Fieldwork Relations’. In: *Contemporary Field Research: A Collection of Readings*. Ed. by R. M. Emerson. Prospect Heights, IL: Waveland Press, pp. 235–252.
- Postill, John (2010). ‘Introduction: Theorising Media and Practice’. In: *Theorising Media and Practice*. Ed. by Birgit Bräuchler and John Postill. New York: Berghahn Books.
- Prasad, P. (1997). ‘Systems of Meaning: Ethnography as a Methodology for the Study of Information Technologies’. In: *Information Systems and Qualitative Research: Proceedings of the IFIP TC8 WG 8.2 International Conference on Information Systems and Qualitative Research, 31st May - 3rd June 1997, Philadelphia, Pennsylvania, USA*. Ed. by Allen S. Lee, Jonathan Liebenau and Janice I. DeGross. Boston, MA: Springer US, pp. 101–118. URL: http://dx.doi.org/10.1007/978-0-387-35309-8_7.
- Prior, Lindsay (2003). *Using Documents in Social Research*. London: SAGE.

- Quinn, Michelle (2019). *How Silicon Valley is finally growing up (sort of)*. Magazine. URL: <https://www.nationalgeographic.com/magazine/2019/02/silicon-valley-evolving-focusing-employees/> (visited on 12th Jan. 2019).
- Ramesh, Randeep and Alex Hern (2013). 'Conservative party deletes archive of speeches from internet'. In: *The Guardian*. ISSN: 0261-3077. URL: <https://www.theguardian.com/politics/2013/nov/13/conservative-party-archive-speeches-internet> (visited on 10th July 2019).
- Raymond, Matt (2010). *How Tweet It Is! Library Acquires Entire Twitter Archive*. Blog. URL: <http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/> (visited on 15th Apr. 2016).
- Reid, Elizabeth (1996). 'Communication and Community on Internet Relay Chat: Constructing Communities'. In: *High Noon on the Electronic Frontier: Conceptual Issues in Cyberspace*. Ed. by Peter Ludlow. Cambridge, MA: MIT Press, pp. 397–412.
- Reissman, Catherine Kohler (1993). *Narrative Analysis*. Vol. 30. Qualitative Research Methods. London: SAGE.
- Renninger, Bryce J (2015). "Where I can be myself ... where I can speak my mind" : Networked counterpublics in a polymedia environment'. In: *New Media & Society* 17.9, pp. 1513–1529. ISSN: 1461-4448. URL: <https://doi.org/10.1177/1461444814530095> (visited on 26th Feb. 2019).
- Reyes Ayala, Brenda (2018). *A Grounded Theory of Information Quality for Web Archives*. Dissertation Defense. University of North Texas College of Information. URL: <https://digital.library.unt.edu/ark:/67531/metadc1181153> (visited on 28th June 2018).
- Ribes, David and Steven J. Jackson (2013). 'Data Bite Man: The Work of Sustaining a Long-Term Study'. In: *"Raw Data" is an Oxymoron*. Ed. by Lisa Gitelman. Cambridge, MA; London, England: MIT Press, pp. 147–166.
- Ridener, John (2009). *From Polders to Postmodernism: A Concise History of Archival Theory*. Duluth, MN, USA: Litwin Books.
- Rinberg, Toly et al. (2018). *Changing the Digital Climate: How Climate Change Web Content is Being Censored Under the Trump Administration*. Environmental Data & Governance Initiative. URL: <https://enviroidatagov.org/wp-content/uploads/2018/01/Part-3-Changing-the-Digital-Climate.pdf> (visited on 31st July 2019).
- Rinehart, David (2016). *Internet Archive 20th Anniversary Event Images*. Internet Archive. URL: https://archive.org/stream/ia20thanniversaryevent_images/ia20thanniversaryevent-rinehart (visited on 13th June 2019).
- Rogers, Richard (2004). *Information Politics on the Web*. Cambridge, Massachusetts: MIT Press.
- (2013). *Digital Methods*. Cambridge, MA: MIT Press.
- (2014). 'Debanalising Twitter: The Transformation of an Object of Study'. In: *Twitter and Society*. Ed. by Katrin Weller et al. Digital Formations. New York, NY: Peter Lang, pp. ix–xxiii.

- (2017). ‘Doing Web history with the Internet Archive: screencast documentaries’. In: *Internet Histories* 1.1, pp. 160–172. ISSN: 2470-1475. DOI: 10.1080/24701475.2017.1307542. URL: <https://doi.org/10.1080/24701475.2017.1307542> (visited on 14th July 2019).
- Rooke, Alison (2009). ‘Queer in the Field: On the messy matters of ethnographic research’. In: *Journal of Lesbian Studies* 13.2, pp. 149–160.
- Rosenthal, David (2016). *Fixing broken links in Wikipedia*. Blog. URL: <http://blog.dshr.org/2016/11/fixing-broken-links-in-wikipedia.html> (visited on 5th Feb. 2017).
- Rosenthal, Uriel, Michael T. Charles and Paul ‘t Hart (1989). *Coping with Crises: The Management of Disasters, Riots and Terrorism*. Springfield, Illinois: Charles C. Thomas.
- Rossi, Alexi (2017). *If You See Something, Save Something - 6 Ways to Save Pages In the Wayback Machine*. Internet Archive Blogs. URL: <https://blog.archive.org/2017/01/25/see-something-save-something/> (visited on 10th July 2019).
- Ruest, Nick and Ian Milligan (2016). ‘An Open-Source Strategy for Documenting Events: The Case Study of the 42nd Canadian Federal Election on Twitter’. In: *Code4Lib* 32. URL: <http://journal.code4lib.org/articles/11358> (visited on 26th Apr. 2016).
- Ryle, Gilbert (1968). *The Thinking of Thoughts: What is ‘Le Penseur’ Doing?* University Lectures No. 18. Saskatchewan: University of Saskatchewan. URL: https://web.archive.org/web/20141221022028/http://lucy.ukc.ac.uk/CSACSIA/Vol114/Papers/ryle_1.html (visited on 19th July 2019).
- Said, Carolyn (1998). *Archiving the Internet / Brewster Kahle makes digital snapshots of Web*. SFGate. URL: <https://www.sfgate.com/business/article/Archiving-the-Internet-Brewster-Kahle-makes-3006888.php> (visited on 25th July 2019).
- SalahEldeen, Hany M. and Michael L. Nelson (2012). ‘Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost?’ In: *Theory and Practice of Digital Libraries: Second International Conference, TPDL 2012, Paphos, Cyprus, September 23-27, 2012. Proceedings*. Ed. by Panayiotis Zaphiris et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 125–137. URL: http://dx.doi.org/10.1007/978-3-642-33290-6_14.
- Sandvig, Christian and Eszter Hargittai (2015). ‘How to Think about Digital Research’. In: *Digital Research Confidential: The Secrets of Studying Behavior Online*. Ed. by Eszter Hargittai and Christian Sandvig. Cambridge, Massachusetts; London, England: MIT Press, pp. 1–28.
- Schatzki, Theodore R. (2001). ‘Introduction: practice theory’. In: *The Practice Turn in Contemporary Theory*. Ed. by Theodore R. Schatzki, Karin Knorr Cetina and Eike von Savigny. London and New York: Routledge, pp. 10–23.
- Schatzman, Leonard and Anselm Strauss (1973). *Field Research: Strategies for a Natural Sociology*. Prentice-Hall Methods of Social Science Series. Englewood Cliffs, NJ: Prentice-Hall.

- Schneider, Steve and Kirsten Foot (2008). 'Archiving of Internet Content'. In: *The International Encyclopedia of Communication*. Ed. by Wolfgang Donsbach. Wiley Publishing.
- Schneider, Steven M. and Kirsten A. Foot (2002). 'Online Structure for Political Action: Exploring Presidential Campaign Web Sites from the 2000 American Election'. In: *Javnost - The Public* 9.2, pp. 43–59. ISSN: 1318-3222. DOI: 10.1080/13183222.2002.11008799. URL: <https://doi.org/10.1080/13183222.2002.11008799> (visited on 1st Aug. 2019).
- (2004). 'The Web as an Object of Study'. In: *New Media & Society* 6.1, pp. 114–122. URL: <https://doi.org/10.1177/1461444804039912>.
- Schneider, Steven M., Kirsten A. Foot and Paul Wouters (2009). 'Web Archiving as e-Research'. In: *e-Research: Transformation in Scholarly Practice*. Ed. by Nicholas W. Jankowski. Routledge.
- Schostag, Sabine and Eva Fønss-Jørgensen (2012). 'Webarchiving: Legal Deposit of Internet in Denmark. A Curatorial Perspective'. In: *Microform & Digitization Review* 41.3-4, pp. 110–120. URL: <http://dx.doi.org/10.1515/mir-2012-0018> (visited on 14th Apr. 2016).
- Schroeder, Ralph and Niels Brügger (2017). 'Introduction: The web as history'. In: *The Web as History: Using Web Archives to Understand the Past and the Present*. Ed. by Niels Brügger and Ralph Schroeder. London: UCL Press, pp. 1–19.
- Schwartz, Joan M. and Terry Cook (2002). 'Archives, Records, and Power: The Making of Modern Memory'. In: *Archival Science* 2, pp. 1–19.
- Scott, Jason (2008). *Eviction, or the Coming Datapocalypse*. ASCII by Jason Scott. URL: <http://ascii.textfiles.com/archives/1617> (visited on 3rd July 2019).
- (2009a). *Datapocalypse!* ASCII by Jason Scott. URL: <http://ascii.textfiles.com/archives/1649> (visited on 10th Jan. 2017).
- (2009b). *The Continuing Adventure of Archive Team*. Blog. URL: <http://ascii.textfiles.com/archives/1886> (visited on 10th Jan. 2017).
- (2011). *Archive Team: A Distributed Preservation of Service Attack*. Conference. Las Vegas, Nevada. URL: <https://www.youtube.com/watch?v=-2ZTmuX3cog> (visited on 10th Jan. 2017).
- (2012). *Open Source, Open Hostility, Open Doors*. Conference Presentation. Portland, Oregon, USA. URL: <https://youtu.be/tJqZGRIwtXk> (visited on 15th Jan. 2019).
- (2017). *Robots.txt is a suicide note*. Archive Team Wiki. URL: <https://www.archiveteam.org/index.php?title=Robots.txt> (visited on 28th July 2019).
- Seaver, Nick (2017). 'Algorithms as culture: Some tactics for the ethnography of algorithmic systems'. In: *Big Data & Society* 4.2. DOI: 10.1177/2053951717738104. URL: <https://doi.org/10.1177/2053951717738104>.
- Seko, Yukari and Stephen P Lewis (2018). 'The self-harmed, visualized, and reblogged: Remaking of self-injury narratives on Tumblr'. In: *New Media & Society* 20.1,

- pp. 180–198. ISSN: 1461-4448. URL: <https://doi.org/10.1177/1461444816660783> (visited on 2nd Feb. 2019).
- Sellers, Christopher et al. (2017). *The EPA Under Siege: 100 Days and Counting*. Environmental Data & Governance Initiative. URL: <https://envirodatagov.org/publication/the-epa-under-siege> (visited on 10th Jan. 2018).
- Sewell Jr., William H. (1999). 'The Concept(s) of Culture'. In: *Practicing History: New Directions in Historical Writing after the Linguistic Turn*. London: Taylor and Francis, pp. 35–61. URL: <https://doi.org/10.4324/9780203335697> (visited on 29th June 2019).
- Shankar, Kalpana (2004). 'Recordkeeping in the Production of Scientific Knowledge: An Ethnographic Study'. In: *Archival Science* 4, pp. 367–382.
- Shapin, Steven (1998). 'Placing the View from Nowhere: Historical and Sociological Problems in the Location of Science'. In: *Transactions of the Institute of British Geographers* 23.1, pp. 5–12. ISSN: 0020-2754. DOI: 10.1111/j.0020-2754.1998.00005.x. URL: <https://rgs-ibg.onlinelibrary.wiley.com/doi/abs/10.1111/j.0020-2754.1998.00005.x> (visited on 29th July 2019).
- Shechmeister, Matthew (2009). 'Ghost Pages: A Wired.com Farewell to GeoCities'. In: *Wired*. ISSN: 1059-1028. URL: <https://www.wired.com/2009/11/geocities/> (visited on 14th June 2020).
- Spaniol, Marc et al. (2009). 'Data Quality in Web Archiving'. In: *Proceedings of the 3rd Workshop on Information Credibility on the Web*. New York, NY, USA: ACM, pp. 19–26. URL: <http://doi.acm.org/10.1145/1526993.1526999> (visited on 11th Jan. 2017).
- Spradley, James P. (1979). *The Ethnographic Interview*. United States: Holt, Rinehart and Winston.
- (1980). *Participant Observation*. United States: Wadsworth/Thomson Learning.
- Spring Owl Asset Management LLC (2015). *Yahoo! Investor Presentation: A Better Plan For Yahoo Shareholders*. Shareholders. URL: <http://web.archive.org/web/20190201233448/http://www.wsj.com/public/resources/documents/yahoopresentation.pdf> (visited on 1st Feb. 2019).
- Srinivasan, Venkat (2016). *The Internet Archive – Bricks and Mortar Version*. Blog. URL: <https://blogs.scientificamerican.com/guest-blog/the-internet-archive-bricks-and-mortar-version/> (visited on 4th Feb. 2017).
- Star, Susan Leigh (1999). 'The Ethnography of Infrastructure'. In: *American Behavioral Scientist* 43.3, pp. 377–391. URL: <http://dx.doi.org/10.1177/00027649921955326>.
- Star, Susan Leigh and James R. Griesemer (2015). 'Institutional Ecology, "Translations," and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-1939'. In: *Boundary Objects and Beyond: Working with Leigh Star*. Ed. by Geoffrey C. Bowker et al. Cambridge, MA; London, England: MIT Press, pp. 171–200.

- Star, Susan Leigh and Karen Ruhleder (1996). 'Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces'. In: *Information Systems Research* 7.1.
- Stefik, Mark (1997). 'The Digital Library Metaphor: The I-Way as Publishing and Community Memory'. In: *Internet Dreams: Archetypes, Myths, and Metaphors*. Ed. by Mark Stefik. Paperback. Cambridge, MA; London, England: MIT Press, pp. 1–14.
- Sterne, Jonathan (2014). "What Do Web Want?" "Materiality!" "When Do We Want It?" "Now!" In: *Media Technologies: Essays on Communication, Materiality, and Society*. Ed. by Tarleton Gillespie, Pablo J. Boczkowski and Kirsten A. Foot. Cambridge, Massachusetts; London, England: MIT Press, pp. 119–128.
- Stoler, Ann Laura (2002). 'Colonial archives and the arts of governance.' In: *Archival Science* 2.1-2, pp. 87–109. URL: <http://rd.springer.com/article/10.1007%2F02435632> (visited on 3rd Sept. 2015).
- Strauss, Anselm (1988). 'The Articulation of Project Work: An Organizational Process'. en. In: *The Sociological Quarterly* 29.2, pp. 163–178. ISSN: 1533-8525. URL: <https://doi.org/10.1111/j.1533-8525.1988.tb01249.x> (visited on 13th May 2019).
- Suchman, Lucy A. (1985). *Plans and Situated Actions: The problem of human-machine communication*. Tech. rep. ISL-6. Palo Alto, CA: Xerox. URL: http://bitsavers.trailing-edge.com/pdf/xerox/parc/techReports/ISL-6_Plans_and_Situated_Actions.pdf (visited on 7th Sept. 2014).
- (2001). 'Building Bridges: Practice-based Ethnographies of Contemporary Technology'. In: *Anthropological Perspectives on Technology*. Ed. by Michael Schiffer. Albuquerque: University of New Mexico Press, pp. 163–177.
- Summers, Ed (2017). *robots.txt*. Blog. URL: <https://inkdroid.org/2017/04/23/robots/> (visited on 25th Feb. 2019).
- Summers, Ed and Ricardo Punzalan (2017). 'Bots, Seeds and People: Web Archives As Infrastructure'. In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '17. New York, NY, USA: ACM, pp. 821–834. URL: <http://doi.acm.org/10.1145/2998181.2998345>.
- Swidler, Ann (1986). 'Culture in Action: Symbols and Strategies'. In: *American Sociological Review* 51.2, pp. 273–286. ISSN: 0003-1224. DOI: 10.2307/2095521. URL: <https://www.jstor.org/stable/2095521> (visited on 29th June 2019).
- (2001). 'What anchors cultural practices'. In: *The Practice Turn in Contemporary Theory*. Ed. by Theodore R. Schatzki, Karin Knorr Cetina and Eike von Savigny. London and New York: Routledge, pp. 83–102.
- Swisher, Kara (2017). *Full transcript: Internet Archive founder Brewster Kahle on Re-code Decode*. In collab. with Brewster Kahle. URL: <https://www.recode.net/2017/3/8/14843408/transcript-internet-archive-founder-brewster-kahle-wayback-machine-recode-decode> (visited on 7th Apr. 2019).
- Taylor, Diana (2003). *The Archive and the Repertoire: Performing Cultural Memory in the Americas*. London: Duke University Press.

- Taylor, Gary W. and Jane M. Ussher (2001). 'Making Sense of S&M: A Discourse Analytic Account'. In: *Sexualities* 4.3, pp. 293–314. ISSN: 1363-4607. DOI: 10.1177/136346001004003002. URL: <https://doi.org/10.1177/136346001004003002> (visited on 22nd July 2019).
- Taylor, Nicholas (2017). 'Introduction to the Special Issue on Web Archiving'. In: *Journal of Western Archives* 8.2. URL: <http://digitalcommons.usu.edu/westernarchives/vol8/iss2/1/>.
- Theimer, Kate (2012). 'Archives in Context and as Context Journal of Digital Humanities'. In: *Journal of Digital Humanities* 1.2. URL: <http://journalofdigitalhumanities.org/1-2/archives-in-context-and-as-context-by-kate-theimer/> (visited on 21st June 2020).
- Thelwall, Mike and David Stuart (2006). 'Web crawling ethics revisited: Cost, privacy, and denial of service'. In: *Journal of the American Society for Information Science and Technology* 57.13, pp. 1771–1779. URL: <http://dx.doi.org/10.1002/asi.20388>.
- Thelwall, Mike and Liwen Vaughan (2004). 'A fair history of the Web? Examining country balance in the Internet Archive'. In: *Library & Information Science Research* 26, pp. 162–176.
- Thrift, Nigel (2005). *Knowing Capitalism*. London; Thousand Oaks, New Delhi: SAGE.
- Tiidenberg, Katrin (2014). 'There's no limit to your love - scripting the polyamorous self.' In: *Journal für Psychologie* 22.1, pp. 1–27. URL: <https://www.journal-fuer-psychologie.de/index.php/jfp/article/view/320> (visited on 2nd Feb. 2019).
- (2016). 'Boundaries and conflict in a NSFW community on tumblr: The meanings and uses of selfies'. In: *New Media & Society* 18.8, pp. 1563–1578. URL: <https://doi.org/10.1177/1461444814567984>.
- Trace, Ciaran B. (2002). 'What is recorded is never simply 'what happened': Record keeping in modern organizational culture'. In: *Archival Science* 2.1, pp. 137–159. URL: <http://dx.doi.org/10.1007/BF02435634>.
- Traweek, Sharon (1992). *Beamtimes and Lifetimes: The World of High Energy Physicists*. Paperback. Cambridge, Massachusetts; London, England: Harvard University Press.
- Trouillot, Michel-Rolph (1995). *Silencing the Past: Power and the Production of History*. Boston: Beacon Press.
- Truman, Gail (2016). *Web Archiving Environmental Scan*. Tech. rep. Cambridge, MA: Harvard Library. URL: <http://nrs.harvard.edu/urn-3:HUL.InstRepos:25658314> (visited on 3rd May 2016).
- Turnbull, David (2003). 'Assemblages and diversity: Working with incomensurability: Emergent knowledge, narrativity, performativity, mobility and synergy'. In: Melbourne. URL: <http://thoughtmesh.net/publish/printable.php?id=279> (visited on 11th Mar. 2016).
- Ubois, Jeff (2016). *An Intimate Conversation with Brewster Kahle - Library Leaders Forum 2016*. In collab. with Brewster Kahle.

- UK Data Forum (2013). *UK Strategy for Data Resources for Social and Economic Research 2013-2018*. Tech. rep. ESRC UK Data Forum. URL: <http://www.esrc.ac.uk/files/research/uk-strategy-for-data-resources-for-social-and-economic-research/> (visited on 8th Apr. 2016).
- UK Parliament (2013). *Legal Deposit Libraries (Non-Print Works) Regulations*. URL: http://www.legislation.gov.uk/ukxi/2013/777/pdfs/ukxi_20130777_en.pdf.
- Uprichard, Emma (2012). 'Being stuck in (live) time: the sticky sociological imagination'. In: *The Sociological Review* 60, pp. 124–138. URL: <http://dx.doi.org/10.1111/j.1467-954X.2012.002120.x>.
- Upward, Frank (1997). 'Structuring the records continuum part two. Structuration theory and recordkeeping'. In: *Archives & Manuscripts*, pp. 10–35. URL: <https://publications.archivists.org.au/index.php/asa/article/view/8613> (visited on 28th June 2020).
- USGAO (2015). *Library of Congress: Strong Leadership Needed to Address Serious Information Technology Management Weaknesses*. Report to Congressional Committees GAO-15-315. Washington, D.C.: United States Government Accountability Office, pp. 1–128. URL: <http://www.gao.gov/assets/670/669367.pdf>.
- Vera, Lourdes A. et al. (2018). 'Data resistance: a social movement organizational autoethnography of the environmental data and governance initiative'. In: *Mobilization: An International Quarterly* 23.4, pp. 511–529. ISSN: 1086-671X. URL: <https://mobilizationjournal.org/doi/abs/10.17813/1086-671X-24-4-511> (visited on 24th Apr. 2019).
- Vine (2016). *Important News about Vine*. Blog. URL: <https://medium.com/@vine/important-news-about-vine-909c5f4ae7a7#.aa1grwtal> (visited on 7th Jan. 2017).
- Vlassenroot, Eveline et al. (2019). 'Web archives as a data resource for digital scholars'. In: *International Journal of Digital Humanities*. ISSN: 2524-7840. URL: <https://doi.org/10.1007/s42803-019-00007-7>.
- Walker, Dawn (2017a). 'Ensuring Climate Data Remains Public'. Presentation at the 34th Chaos Communication Congress. URL: <https://doi.org/10.5446/34854> (visited on 11th May 2019).
- Walker, Dawn et al. (2018). 'Practicing environmental data justice: From DataRescue to Data Together'. In: *Geo: Geography and Environment* 5.2, pp. 1–14. ISSN: 2054-4049. URL: <https://doi.org/10.1002/geo2.61>.
- Walker, Shawn (2017b). 'The Complexity of Collecting Digital and Social Media Data in Ephemeral Contexts'. PhD. University of Washington. URL: <http://hdl.handle.net/1773/40612> (visited on 26th Nov. 2018).
- Wallace, Gregory (2018). *Report: Trump admin scrubbed mentions of climate change from websites*. CNN. URL: <https://www.cnn.com/2018/01/10/politics/trump-climate-change-websites-remove-report/index.html> (visited on 9th July 2019).

- Warde, A (2005). 'Consumption and Theories of Practice'. In: *Journal of Consumer Culture* 5, pp. 131–153.
- Waterton, Claire (2010). 'Experimenting with the Archive: STS-ers As Analysts and Co-constructors of Databases and Other Archival Forms'. In: *Science, Technology, & Human Values* 35.5, pp. 645–676.
- Watson, Kathryn (2018). 'White House updates online transcript of Trump-Putin news conference'. In: *CBS News*. URL: <https://www.cbsnews.com/news/white-house-updates-online-transcript-of-trump-putin-news-conference/> (visited on 13th Mar. 2019).
- Webb, Collin, David Pearson and Paul Koerbin (2013). "Oh, you wanted us to preserve that?!" Statements of Preservation Intent for the National Library of Australia's Digital Collections'. In: *D-Lib Magazine* 19.1/2. URL: <http://www.dlib.org/dlib/january13/webb/01webb.print.html> (visited on 30th Jan. 2016).
- Weber, Matthew S. and Philip M. Napoli (2018). 'Journalism History, Web Archives, and New Methods for Understanding the Evolution of Digital Journalism'. In: *Digital Journalism*, pp. 1–20. ISSN: 2167-0811, 2167-082X. DOI: 10.1080/21670811.2018.1510293. URL: <https://www.tandfonline.com/doi/full/10.1080/21670811.2018.1510293> (visited on 11th Oct. 2018).
- Webster, Peter (2013). *Political party web archives*. UK Web Archive blog. URL: <https://blogs.bl.uk/webarchive/2013/12/political-party-web-archives.html> (visited on 9th July 2019).
- (2017). 'Users, technologies, organisations: Towards a cultural history of world web archiving'. In: *Web 25: Histories from the First 25 Years of the World Wide Web*. Ed. by Niels Brügger. Peter Lang, pp. 170–190.
- Winner, Langdon (1993). 'Upon Opening the Black Box and Finding It Empty: Social Constructivism and the Philosophy of Technology'. In: *Science, Technology, & Human Values* 18.3, pp. 362–378. URL: <http://www.jstor.org/stable/689726>.
- Winters, Jane. 'Giving with one hand, taking with the other: e-legal deposit, web archives and researcher access'. In: *Electronic Legal Deposit: Shaping the library collections of the future*. Ed. by Paul Gooding and Melissa Terras. London: Facet Publishing.
- Wittel, Andreas (2001). 'Toward a Network Sociality'. In: *Theory, Culture & Society* 18.6, pp. 51–76. URL: <https://doi.org/10.1177/026327601018006003>.
- Yakel, Elizabeth (1997). 'Record-keeping in Radiology: The Relationship Between Activities and Records in Radiological Processes'. PhD. Ann Arbor, Michigan: University of Michigan.
- (2001). 'The Social Construction of Accountability: Radiologists and Their Record-Keeping Practices'. In: *The Information Society* 17.4, pp. 233–245. URL: <http://dx.doi.org/10.1080/019722401753330832>.
- Yin, Robert K. (2009). *Case Study Research: Design and Methods*. Fourth. Applied Social Science Research Methods Series 5. London: SAGE.

- Yip, Jason (2016). *It's Not Just Standing Up: Patterns for Daily Standup Meetings*. Blog. URL: <https://web.archive.org/web/20181227052106/https://martinfowler.com/articles/itsNotJustStandingUp.html> (visited on 8th Jan. 2019).
- Zannettou, Savvas et al. (2018). 'Understanding Web Archiving Services and Their (Mis)Use on Social Media'. In: *Proceedings of the 12th International AAAI Conference on Web and Social Media*. AAAI Conference on Web and Social Media (ICWSM). Palo Alto, CA: arxiv. URL: <https://arxiv.org/abs/1801.10396> (visited on 9th July 2019).
- Zeitlyn, David (2012). 'Anthropology in and of the Archives: Possible Futures and Contingent Pasts. Archives as Anthropological Surrogates'. In: *Annual Review of Anthropology* 41.1, pp. 461–480. URL: <http://dx.doi.org/10.1146/annurev-anthro-092611-145721> (visited on 13th Mar. 2015).
- Zinn, Howard (1977). 'Secrecy, Archives, and the Public Interest'. In: *The Midwestern Archivist* 2.2, pp. 14–26. URL: <https://www.jstor.org/stable/41101382>.