ABSTRACT

Title of Dissertation: LEGIBILITY MACHINES:
ARCHIVAL APPRAISAL AND THE
GENEALOGIES OF USE

Ed Summers
Doctor of Philosophy, 2020

Dissertation directed by: Associate Professor Ricardo Punzalan
College of Information Studies

The web is a site of constant breakdown in the form of broken links, failed business models, unsustainable infrastructure, obsolescence and general neglect. Some estimate that about a quarter of all links break every 7 years, and even within highly curated regions of the web, such as scholarly publishing, rates of link rot can be as high as 50%. Over the past twenty years web archiving projects at cultural heritage organizations have worked to stem this tide of loss. Yet, we still understand little about the diversity of actors involved in web archiving, and how content is selected for web archives. This is due in large part to the ontological politics of web archives, and how the practice of archiving the web takes place out of sight at the boundaries between human and technical activity.

This dissertation explores appraisal practices in web archives in order to answer two motivating research questions: 1) How is appraisal currently being enacted in web archives? 2) How do definitions of what constitutes a web archive shape the practice of appraisal? In order to answer these questions data was collected from interviews with practicing professionals in web archives, and from a year long ethnographic field study with a large federally funded archive. Method triangulation using thematic analysis, critical discourse analysis and grounded theory generated a thick and layered description of archival practice. The results of this analysis highlight three

fundamental characteristics of appraisal in web archives: time, ontology and use.

The research findings suggest that as expressions of value, appraisal decisions do not simply occur at discrete moments in the life cycle of records. They are instead part of a diverse set of archival processes that repeat and evolve over time. Appraisal in web archives is not bound by a predefined assemblage of actors, technologies and practices. Indeed, artificially limiting our definition of what constitutes a web archive truncates our understanding of how appraisal functions in web archives. Finally, the valuation of web records is inextricably tied to their use in legibility projects, where use is not singular, but part of a genealogy of use, disuse and misuse.

Appraising appraisal along these three axes of time, ontology and use provides insight into the web-memory practices that condition our understanding of the past, and that also work to create our collective present and futures. Explicitly linking appraisal to the many forms of use informs archival studies pedagogy, by establishing the value of records in terms of the processes they participate in, rather than as a static attribute of the records or their immediate context. As machines increasingly become users of web archives the stakes for understanding the values present in web archival platforms could not be higher.

LEGIBILITY MACHINES:
ARCHIVAL APPRAISAL AND THE GENEALOGIES OF USE

by

Ed Summers

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Associate Professor Ricardo Punzalan, Chair
Associate Professor Kari Kraus
Associate Professor Wayne G. Lutters
Assistant Professor Katrina Fenlon
Professor Matthew Kirschenbaum, Dean's Representative

Acknowledgements

One of the difficulties of returning to university to pursue a PhD after decades of working in the field is that there are too many people to thank. I never thought I'd need so many people. Rather than pretend that such a list was actually complete I'd like to thank a few close people who set me on my path, and helped me along the way. Mom and Dad, I never wrote a proper thank you to you in my high school graduation yearbook. I owe so much to you: your teamwork, your care for life and memory, and your hope for the present and the future. I know that the best parts of anything I've done have grown from the love you have given me. Matthew, thank you for always reminding me of the importance of connecting with others, and the value of a quick wit, sense of humor, and the gift of music, when it is most needed. John, thank you for being my intellectual companion, even as the years have separated us in space, I continue to draw on your creative imagination and curiosity. Charlie, Graham and Maeve, thank you for grounding me, for challenging me, and for showing me how to continue to grow as a human being, and as your Dad. Kathy and Ron, thanks for your trust and confidence in me, and showing me how to live for life and family. Finally, Kesa, none of this would have gotten done if it wasn't for you. I know you know this already. I don't say it hardly enough but your joyful light shines so brightly, every day, and helps me see. Thank you for sharing your pragmatism with me: not a cynical ends-justifies-the-means philosophy, but a centering wisdom and concern for the present and the people in it, in all their complexities. Thanks for being here, written between these lines.

The forms that documentary work assumes are as numerous as the needs from which they are born.

Suzanne Briet

Technologies are revolutions in consciousness.

Nikolas Rose

Knowledge lies in exemplars and words are never enough.

John Law

Table of Contents

Acknowledgements i

Table of Contents iii

List of Figures vii

1. Introduction 1

2. Literature Review 7

2.1. Appraisal 7

2.1.1. Archival Meta-theories 8

2.1.2. Governmentality 10

2.1.3. Tacit Appraisal 13

2.1.4. Realizing Appraisal 15

2.1.5. Documenting Society 17

2.1.6. Appraisal Critique 19

2.1.7. Strategy 21

2.1.8. Decentralization 23

2.1.9. Outside the Archive 24

2.1.10. Appraising Appraisal 25

2.2. Appraisal and Web Archives 26

2.2.1. Science and Technology Studies 29

2.2.2. Digital Libraries 30

2.2.3. Electronic Records 32

2.3. Appraisal Practice 35

2.3.1. Ethnography of Infrastructure 38

2.3.2. Repair and Maintenance 39
2.3.3. Software Studies 40
2.3.4. Data Justice 43
2.4. Appraisal 45

3. Methodology **46**
3.1. Algorithms as Culture 48
3.2. Studying Up 50
3.3. Relations All the Way Down 51
3.4. Research Design 53

4. Bots, Seeds and People **58**
4.1. Methodology 59
4.2. Findings 64
4.2.1. Crawl Modalities 65
4.2.2. Information Structures 67
4.2.3. Time and Money 70
4.2.4. People 72
4.2.5. Tools 74
4.2.6. Breakdown 77
4.3. Discussion 80

5. Appraisal Talk in Web Archives **84**
5.1. Methodology 85
5.2. Findings 88
5.2.1. Hierarchies 89
5.2.2. Mentorship 93
5.2.3. Structures 97
5.3. Discussion 100

6. Seeing Software **105**

6.1. Methodology 107

6.2. Findings 114

6.2.1. NIST 114

6.2.2. The Scene 117

6.2.3. Google Play 127

6.2.4. Game Platforms 132

6.2.5. Hashing and Fixity 136

6.2.6. The Customer 149

6.2.7. CAID 156

6.2.8. ByLock 160

6.2.9. Cabrinety 163

6.3. Discussion 171

6.3.1. Use 172

6.3.2. Legibility 175

6.3.3. Governmentality 179

7. A Conclusion **182**

8. Appendixes **194**

A. Study 1 Documents 194

A.1. Consent Form 194

A.2. Recruitment Letter 198

A.3. Interview Protocol 199

B. Study 2 Documents 202

B.1. Consent Form 202

B.2. Recruitment Letter 205

B.3. Interview Protocol 206

9. References **207**

List of Figures

1 Personnel Distribution United States Civil Service Commission . . 17

2 Between-Method Triangulation 56

3 Field Notes and Jottings 110

4 Coding with Anselm 110

5 RDS Field Use 118

6 Administration Building 101 120

7 Chemistry Building 222 121

8 Architecture Sketch: October 10, 2018 127

9 Architecture Sketch: December 13, 2018 127

10 Architecture Sketch: March 3, 2019 128

11 Steam Gift Cards 133

12 The STEAM_APP Table (SQL) 135

13 Staff in the NSRL Library, 2014 138

14 Hash Counts Email 139

15 NSRL Lab Door 142

16 Juggernaut Before Refactoring 146

17 Juggernaut After Refactoring 147

18 The Customer 152

19 NIST (2002) 153

20 CAID Recognition 159

21 CAID and Games Archiving 160

22 Turkish Intelligence Forensics Diagram from Fox-IT (2017) 162

23 Age of Cabrinety Materials 169

24 2013/465/24 Time Curves by Alan Levine 188

25 Concentric Rings of Use 189
26 Age in Double Figures? by Garry Knight 190
27 Multiple Concentric Rings of Use 190

1. Introduction

For a moment try to imagine your day to day life without the web and the underlying Internet. It is difficult to do because the web is a communications infrastructure that is completely enmeshed in global systems of capital, governance, science and culture. However, somewhat paradoxically, we experience the loss of the web on a daily basis, every time we click a link only to get a *404 Not Found* error. The architectural precarity of our "World Wide Web" is constantly being made visible to us. These quotidian breakdowns are so common that we have come to expect, or even anticipate them. The continual loss of the web and its broken links become part of the infrastructural background in the metaphor of the constantly changing cloud (Hu, 2015). But what remains of this backgrounding, or evaporation, is an archival anxiety that asks: What will we remember of our current epoch? Are we really living in what will be known as a *digital dark age*? (Hedstrom, 1991; Kuny, 1997) Or perhaps we are living in the ruins of a digital dark age right now?

Partly in response to this archival anxiety we have witnessed the emergence of *web archiving* as an increasingly significant activity. Web archiving is the practice of collecting content from the web for preservation, which is then made accessible at another part of the web known as a *web archive*. Web archiving is typically preformed by memory institutions (libraries, archives and museums), but increasingly it is individuals who take up the work themselves (C. A. Lee, 2011). Developing record keeping practices for selecting, collecting and preserving web content is proving to be an extremely important endeavor for historical production (Brügger & Schroeder, 2017) and for sustaining the networked public sphere (Benkler, 2006 ; Lessig, 2006). Archivists use the term *appraisal* to talk about the theories and methods that determine what records are collected in an archive. However, even with close to two decades of practice we understand surprisingly little about the day to day processes

by which content is being appraised for web archives.

At the same time, our experience of using the web and the Internet is suffused with constant, and often invisible, data collection. These data flows pool into vast corporate and government data archives that have come to be referred to in shorthand as *Big Data*. For example, in 2015 Facebook was collecting two billion photographs a day from its users, which at that time required hundreds of petabytes of storage (Bandaru & Patiejunas, 2015). Always-on mobile computing technologies, the *Internet of Things* and *smart cities* provide the infrastructure for a host of data capture and processing platforms that have become an essential parts of our daily lives. Hoskins (2018) calls these new data archives, and their attendant processual flows *shadow archives*:

> The archive has traditionally been seen (like other media) as separate and external to the self, as something with institutional status, as variously a place and space for the storage of artefacts of the past that give rise to remembering. Yet, the medial gathering and splintering of individual, social and cultural imaginaries, increasingly networked through sortable and pervasive digital media and communication devices, attach shadow archives to much of everyday life, that also blend and complicate that which was once considered as distinctly public and private. (p. 87)

Shadow archives are physically and conceptually remote, are often beyond our individual control, cognition, and are usually only readable in full by the entities that created them. These archives surface on the web in various ways, but are characterized not by an anxiety about what will be remembered, but rather by what will not be forgotten (Mayer-Schönberger, 2011).

Most importantly for my discussion here, these shadow archives are constructed

both *in* and *of* the web, and operate as web archives just as much as the previously mentioned web archives operated by cultural heritage organizations. They are *shadow* archives in a second sense, in that they sit behind or to the side of normalized conceptions of what *web archives* are, as a specific deployment of software, hardware and networked infrastructure. In addition to manifesting as *Big Data* web archives can take the shape of thematically arranged websites (Fenlon, 2017) or *Small Data* (Abreu & Acker, 2013). Questions of what to collect, what not to collect, what to remember, and what to forget, are sites of controversy and anxiety, that surface on the web, and are entangled with standards, protocols and infrastructure. How is it that our idea of what constitutes a *web archive* have become stabilized, and how does this stabilization relate to our decisions about what to archive?

In this dissertation I explore the art and science of deciding what web archives collect (how they appraise). I suggest that our anxieties around what web archives remember and forget, can be understood by attending to the specific material practices of people working with memory and machines. Web archives are legibility projects, or calculative practices, that include, but also escape normative definitions of web archives as a specific technical apparatus. Recognizing the full shape of web archives through the theoretical lens of *governmentality* (Foucault et al., 2008 ; Lemke, 2019) clarifies how the supposed value of archival records is not a property of the records themselves or their context, but is rather a function of the how the records are embedded in architectures of use over time–their genealogies of use. This insight into web archives also suggests a reorientation in how we think about appraisal in archives more generally.

Deciding what to keep, and what to discard, is a central theme in the field of archival studies–some even say it is the archivist's first and most important responsibility (Cox & Samuels, 1988) or "the most significant archival function" (Brichford, 1977).

Over the past two centuries archival theorists have developed a body of literature around the concept of *appraisal*, which is broadly defined by the Society of American Archivists as the "process of identifying materials offered to an archives that have sufficient value to be accessioned". Document production, which began with the innovation of writing, and rapidly accelerated with the publishing technologies of movable type, the printing press, photography, lithography, xerography, and computer automation has made it increasingly necessary for archivists to recognize their pivotal role in deciding what documents get to be called *archival records*.

As a practical matter, for an archive to exist, appraisal decisions must be made, which necessarily shape the archive over time, and by extension also shape our knowledge of the past (Bearman, 1989 ; Cook, 2011). It is in the particular contingencies of the historical moment that the archive is created, sustained and used (Booms, 1987 ; Harris, 2002). And yet the desire for a technology that will enable a complete archival record of the web, where everything is preserved and remembered, is a strangely persistent idea, or aspirational goal, with many social and political ramifications (Brothman, 2001 ; Mayer-Schönberger, 2011). Reviewing this literature of appraisal, with an eye to understanding the appraisal of content on the web is the first focus of this paper.

Part of the reason for the gap in our understanding about how web content is being selected for preservation is a matter of scale. Considered as a whole the web is an unfathomably large, decentralized and constantly changing information landscape. Unlike a box of photographs in an attic, that may find its way into a Hollinger box on a shelf in an archive, web content seems to come to us from *The Cloud* (Hu, 2015), and appears to resist the archival imagination that has traditionally focused on *information as thing* (Buckland, 1991).

The web is a site of constant breakdown in the form of broken links, failed business

models, unsustainable infrastructure, obsolescence and general neglect. Ceglowski (2011) has estimated that about a quarter of all links break every 7 years. Even within highly curated regions of the web, such as scholarly publishing (Sanderson et al., 2011) and jurisprudence (Zittrain et al., 2014) rates of link rot can be as high as 50%. Web archiving projects work in varying measures to stem this tide of loss: to save what is deemed worth saving before it becomes 404 Not Found. In this light, web archiving can be seen as a form of repair or maintenance work (Graham & Thrift, 2007 ; Steven J. Jackson, 2014) that is conducted by archivists, collaborating with each other, while also deeply engaged with tools and infrastructures and platforms that aid them in their work. Attention to issues of repair and maintenance and the larger field of Science and Technology Studies (STS) suggests an examination of web archiving as a set of material practices that includes activities such as website design, upgrades, storage backups, and the porting of content from one content management system to another. I will review how this lens of repair and maintenance helps us think about web archiving as *practice* forms the second part of my analysis.

The construction of web archives, and the maintenance of the web, entail each other, and present challenges and opportunities for archivists as they work with content creators, computational systems, services and other bespoke software. It is important that our knowledge of these systems be informed by an analysis of the social, technical and material practices by which web content is selected for an archive. Decisions about what to collect from the web in a web archive are *co-produced* by the technical means that are used (Jasanoff, 2006). How appraisal is enacted in web archives is fundamentally both an epistemological and an ontological question. How do web archives generate facts, evidence and knowledge? How does our idea of what constitutes a web archive and the record (Yeo, 2007 , 2008) shape that knowledge? Appraisal practices for the web manifest at the interface layer, which

is itself a fractal of the infrastructure of the web itself.

Considering these topics led me to ask two overarching research questions in this dissertation. First, how is appraisal being enacted in web archives? What do we know about current practices and how web records are valued? Secondly, how do definitions of what constitutes a web archive relate to the practice of appraisal? Relative to archival practice, the web is still fairly new. Understanding the values by which web archives are built requires a clear understanding of what we mean by a *web archive*. I will explore answers to these two questions by analyzing interviews with practicing web archivists and drawing on the results of a year long field study with a large federally funded archive engaged with archiving portions of the web. I will conclude by outlining a research agenda for web archives that opens up from an analysis of appraisal, practice and the web, and links the value of records to their use. To get started it is important to first review what we know about the concept of archival appraisal and especially how it relates to the web.

2. Literature Review

This dissertation aims to advance what we know about the practice of appraisal and selection in the context of web archiving. My examination of this literature shows that appraisal in web archives has been treated both from a social perspective in archival studies, and from a technical perspective in digital library and electronic records research. I highlight how appraisal in web archives is best understood as a sociotechnical practice that does not privilege either social or technical explanations. Foucault's concept of governmentality provides a theoretical lens for examining the diverse set of appraisal "theories" present in archival studies, while recognizing the technical practices of the digital library research community. The lack of studies that investigate the actual practices of appraisal in web archives provides an opening for this dissertation's contribution.

2.1. Appraisal

Generally speaking the field of archival studies is praxis oriented, in that it engages with issues of theory in the context of practice. The specific literature of archival appraisal is a prime example of this praxis orientation. The problem of appraisal is fundamentally concerned with the practical problem of how to select material for an archive given 1) an abundance of records, and 2) a finite amount of resources to store all of them. Cox & Samuels (1988) consider appraisal to be the "first responsibility" of the archivist, and define it broadly as:

> … any selection activity that enables archivists to identify recorded information that has enduring value, primarily for the documentation of modern society (p. 29)

In a recent review of the appraisal literature Anderson (2011) (p. 26) distinguishes between *appraisal* and *selection*, in order to provide clarity about when records

are evaluated (appraisal) versus when they are chosen (selected), and when these activities take place relative to an archive taking custody of the records (accession). However for my purposes here a more expansive, and generalist, view on appraisal is taken, which admits that appraisal is a process by which values are asserted about records. These assertions happen in multiple intersecting timelines and at varying scales that take shape due to repeated, atomic actions of selection. This more general view will be important later when appraisal activities are considered in light of the sociotechnical dimensions of the web.

While archives have existed for millenia (Posner, 1972), it is only over the past several hundred years that archivists have developed the concept of appraisal in order to manage the ever increasing deluge of records, that has been brought upon them, largely by the technologies of record production. Three years before the web was first prototyped by Tim Berners-Lee at CERN, Young (1985) at the Bentley Historical Library reviewed the literature of archival appraisal in the United States and found 178 monographs, articles, reports and unpublished manuals. A search in 2018 for *appraisal* after 1985 in Library and Information Science Source yields some 300 more. A complete and exhaustive survey of this literature while possibly beneficial, does not serve my purpose here, which is to connect the literature of archival appraisal with practices of web archiving. To do this it is useful to examine meta-theories, or conceptual frameworks that have been used to talk about archival appraisal.

2.1.1. Archival Meta-theories

Eastwood (2002) outlines three strains of thinking about archival appraisal that have developed over the past few centuries of archival studies. The first and most commonly held view, is that appraisal is a vehicle for history: we must choose what to remember from the perspective of the future historian who is attempting to under-

stand today (Schellenberg, 1956). The second view is that appraisal (the shaping of the archival record by archivists) is not a task for the archivist because it is the enemy of authenticity, evidence and the ultimately the archive itself (Duranti, 1994; Jenkinson, 1922). The third is that appraisal is an inherently political act, that necessarily carries with it the threat of erasure, while also providing opportunities for social justice, ethical engagement, and as sites for creative record creation (Harris, 2002 ; Punzalan & Caswell, 2016).

Eastwood's characterization of appraisal holds up well today, since it casts the three strands not as an evolution in time, but as a bricolage of approaches that coexist and function simultaneously. However it is important to note that Eastwood discusses these three strands of appraisal against a backdrop of Western democratic societies. Eastwood's thesis is that appraisal practices and theories are put to work in the service of democracy, and are to be understood ultimately as a tool for governance and accountability. This is a theme that I will return too shortly. But before I do that I want to examine another appraisal meta-theory.

In her recent study of appraisal practices Foscarini (2017) draws on the work of Cook (2013) to describe the discursive tensions that exist in the research literature about archival appraisal. Cook identified a general timeline of archival paradigms:

> I want to suggest that since the later nineteenth century, archival identity has shifted, or has been in the process of shifting, through four such paradigms or frameworks or mindsets, as it has struggled, and still struggles, with this memory-evidence tension. I am calling the four frameworks: evidence, memory, identity, and community. It is important to emphasize that these four accumulate across time; they do not entirely replace each other.

Foscarini takes up these four paradigms to examine how they manifest in various

theoretical approaches to appraisal, which largely orient around what can be seen as top-down and bottom-up approaches. In top-down approaches (evidence and memory) the archivist intervenes as little as possible in the service of authenticity and integrity (Jenkinson, 1922), or they analyze organizational hierarchies, structures and activities in an attempt to document society as a whole (Cook, 2004). In the bottom-up approach (identity and community) appraisal is recognized as a creative activity, in which history, memory and social relations are assembled as part of a complex set of activities that are not necessarily centered on an institutional context. Foscarini goes on to note that these ideas about appraisal are not steps along a timeline, but are all still very much with us. For example, much of the last few decades of work in digital preservation have been focused on the design and development of technologies for ensuring authenticity and integrity of data, with the implicit and governing assumption that technology can, or should, allow us to perfectly collect everything.

I am offering my own meta-theory of appraisal specifically for web archives, which relies on Foucault's idea of governmentality as a framework for understanding the full scope of appraisal theories, in their social and historical contexts (Foucault, 1991). So before diving into these details it is important to first take a look at *governmentality*, which we will return to later as an example of a practice orientation to web archives.

2.1.2. Governmentality

The records from our earliest archives happen to be coterminous with the beginnings of recorded human history. This is no coincidence given that the methods of history depend in large part on the evidence of the past that survives, and this evidence is often found in archival repositories of various kinds. As such, archives are often seen as historiographical instruments. The disciplines of history and archival stud-

ies are twinned, but their relations are somewhat antagonistic. Historians are one of the most active users of archives. The records collected in archives are used by historians as evidence of previous events and activities, which are essential for piecing together and anchoring historical narratives in fact, or as close as can be got to fact many years later. Until the professionalization of archival studies, it was common for archivists to be trained historians, with invested knowledge of the events, people and organizations that the archive is ostensibly concerned with. Who better to tend to the records of an organization than the scholar who knows its history? It wasn't until archivists began to meaningfully grapple with the concept of *appraisal* that an understanding of the archive developed that was more than simply the tool of memory and history (Taylor, 1984), but also as an political and economic instrument of power, accountability, resistance and liberation (Jimerson, 2009).

The archival practices of southern Mesopotamia and ancient Egypt, which archival studies points to for its origin story (at least in Western traditions), are examples from the earliest known states used archives to manage records of laws, administrative activities, financial transactions, land ownership, and taxation (O'Toole, 2004; Posner, 1972). These archives reflected the interests of the earliest states in governing and controlling their populations. Similarly the gaps in records, and their silences, work to demarcate those on the periphery of the state, or outside it altogether (Scott, 2017). A significant portion of the stories of archives are thus bound up with the interests of institutions, states, governance and what Foucault (1991) calls *governmentality*.

For Foucault governmentality is a mode of rationality of governing through technologies of power, that reproduce themselves through specific practices. Governmentality is not simply the story of the operations of the state, but is concerned with "the conduct of conduct", and can be found in the practices of individuals, house-

holds, families and communities. Foucault traces the emergence of the art of government and political economy, and situates it with the decline of the sovereign, the birth of the modern state, and more recently, neo-liberalism. Governmentality is a useful instrument because it provides a continuous field that includes many modes of archival production and appraisal. It provides a frame for looking at appraisal landscapes as a form of knowledge (savoir), that has particular rules of formation, transformation and correlation with other practices. As noted by Schwartz & Cook (2002):

> Archives have always been about power, whether it is the power of the state, the church, the corporation, the family, the public, or the individual. Archives have the power to privilege and to marginalize. They can be a tool of hegemony; they can be a tool of resistance. They both reflect and constitute power relations. They are a product of society's need for information, and the abundance and circulation of documents reflects the importance placed on information in society. They are the basis for and validation of the stories we tell ourselves, the story-telling narratives that give cohesion and meaning to individuals, groups and societies. (p. 13)

Foucault's idea of governmentality is useful for tracing practices of archival appraisal, or the practices of deciding what to remember and forget, because of the insights it brings into how appraisal practices function as part of the sociotechnical assemblages of archives and their manifestations in the web. But before turning to look at web archives I want to briefly outline a genealogy of appraisal practices that foregrounds governmentality.

2.1.3. Tacit Appraisal

Despite the millennial historical arc of archival practice, Western archival studies often locates its theoretical origins in the concept of *respect des fonds*, or more commonly, *provenance* from 19th century France. Respect des fonds is a practice of grouping records by their creator, rather than by subject matter or some other predetermined taxonomy. In practice this means records are grouped together by the organization, family, or individual that created them. As Bailey (2013) describes, the introduction of provenance occurred in the aftermath of the French Revolution, which saw the simultaneous destruction and reconstruction of records; a truly monumental attempt to both erase and refashion the historical record in the newly established Archives Nationales. While attempting to deal with this project the librarian, archivist and historian Natalis de Wailly introduced the idea of *respect des fonds* in 1841 as a simplified practice for arranging the records:

> The principle was, in fact, a practical exigency, a method to simplify archival arrangement seen as more easily implemented by novice archivists than the more rigorous arrangement by classification. Bailey (2013)

It is significant to note that this use of provenance as a heuristic for organizing records was born amidst what was effectively a massive appraisal process, in which overtly political interests determined the preservation of pre and post-revolution records. However, the concept of appraisal was not itself explicitly part of the expression of respect des fonds. The question of what to keep and what to discard was subsumed into a practice for processing the pre-revolutionary materials.

Despite its partial application in France, respect des fonds proved popular as it spread to Belgium and the Netherlands, where it became part of the *Manual for the Arrangement and Description of Archives* (Muller et al., 1940), that was orig-

inally published in 1898. The so called *Dutch Manual* added to to the concept of respect des fonds an additional principle named *original order*, which stated that the arrangement of records should be the same as the original organization. In their manual the authors Muller, Feith and Fruin describe how records are created during the activities of administrative bodies or officials that operate using the metaphor of a life-form:

> … an archival collection is an organic whole, a living organism which grows, takes shape, and undergoes changes in accordance with fixed rules. If the functions of the body change, the nature of the archival collection changes likewise. The rules which govern the composition, the arrangement and the formation of an archival collection, therefore, cannot be fixed by the archivist in advance; he can only study the organism and ascertain the rules under which it was formed. (p. 19)

Even here the idea of appraisal, or the decision of what to keep and what to discard, is tacit. The concept of original order works to prevent the reordering or disruption of records that have *already* arrived as *archival*. The decision of what administrative bodies and officials to collect has already been made, and the principle of original order works to govern how the records are to be arranged and described once they arrive. The surprisingly fresh conception of records as both product and part of a form of life (organicity), still works to bracket off appraisal as something that has already happened out of band. The fact that an administrative body is generating the records is enough to transmute the documents from mere papers into *archival records* that are to be preserved for the long term. The process of governance is at work in the very conception of how the archive functions.

2.1.4. Realizing Appraisal

In the early 20th century, Hillary Jenkinson imported the concept of respect des fonds from continental Europe and fused it with existing English archival practices in his influential *Manual of Archival Administration*. Jenkinson was a medievalist by training, and stressed the importance of impartiality, authenticity and naturalness (similar to organicity) in archival practice. These three principles coordinate to position the archivist as the *keeper* of records, and proscribes the archivist from acting in any way to shape was is made archival. While Jenkinson admits that decisions need to be made about what to keep, these decisions are made by the record creator, before the records are added to an archive, and not by the archivist after the fact. More recently Duranti (1994) connects ideas around authenticity to the theory of *diplomatics*, the critical and forensic analysis of documents, which she traces back to the practice of Roman law in the 11th century. Notice here that the locus of appraisal is still in the hands of officials working within administrative bodies, in the bureaucracies of power.

Tschan (2002) suggests that much of the last century of archival thinking can be characterized as a sustained conversation between Jenkinson's Manual on the one hand, where archival appraisal is verboten and authenticity is paramount, and another manual *Modern Archives: Principles and Techniques* written by historian and archivist Theodore Schellenberg. Schellenberg's manual was written after World War Two, in a moment when record production was vastly outstripping the ability to store them. In his position at the US National Archives and Records Administration, Schellenberg recognized that pragmatic decisions needed to be made about what records to make *archival*, and that those decisions were essentially assertions about value, of which there were two kinds: primary and secondary.

Primary value was the value of the records to the record creator, in their active use.

Secondary value on the other hand divided into two subtypes: *evidential* and *informational*. On the surface *evidential value* seems to be quite similar in principle to Jenkinson's interest in impartiality, authenticity and naturalness. However Schellberg makes clear that he is not concerned with the fidelity of the records as evidence, but with how well the records function as evidence of the organization and function of a particular government body:

> By evidential value I do not refer here to the value that inheres in public records because of the merit of the evidence they contain. I do not refer, in a Jenkinsonian sense, to the sanctity of the evidence in archives that is derived from "unbroken custody." I refer rather, and quite arbitrarily, to a value that depends on the importance of the matter evidenced, i.e. the organization and functioning of the agency that produced the records.

Schellenberg outlined a variety of criteria to use for assessing the evidentiary value of records, which crucially links the volume of records with organizational hierarchy:

In contrast, *information value* is determined by the archivist in a subjective way, that takes into account the historical moment that the records were created in, and often involves outside consultation with relevant subject matter specialists. Schellenberg, himself trained as a historian, notes that many archivists themselves are historians and thus are "competent to ascertain the historical values of public records" (Schellenberg, 1956, p. 150). As opposed to a highly structured approach, Schellenberg stressed that determinations of *informational value* by archivists resist consistency and systematization. As a result measures of *informational value* will be different in different contexts, and that "diverse judgments may well assure a more adequate social documentation".

Even with these allowances for contextual variation, Schellenberg's notions of ev-

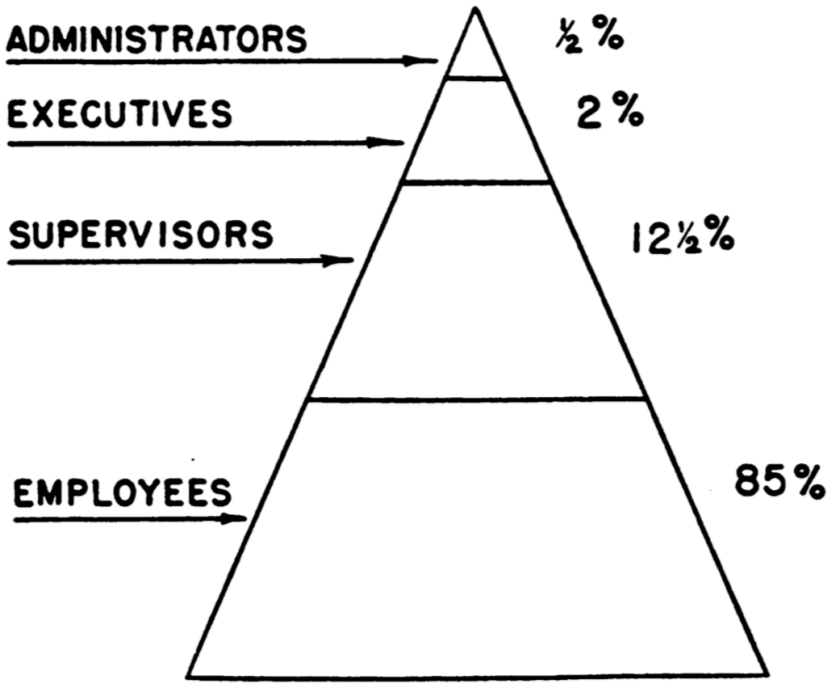

Figure 1: Personnel Distribution United States Civil Service Commission

idential and informational are significantly framed by the institutional and governmental context. This is plainly seen in his use of the hierarchy of government as a mode of selection in Figure 1. In addition, judgments about historical significance made by archivists and other subject specialists, are key to establishing information value, but are themselves bound up and subject to less visible professional structures of power, which are left largely unexamined. Into this gap we see the final stage in top-down approaches, as the unit of analysis moves from the level of the state up yet further to that of society at large.

2.1.5. Documenting Society

The archival theorist Terry Cook was an eloquent archival theorist, most remembered for his ability to synthesize archival theory, and mobilize it for work in the present, particularly in the service of appraisal. His theory of *macroappraisal* provides a framework for shifting focus away from the value of records, and towards

the values inherent in the functional context that records are created in (Cook, 2005). When using macroappraisal archivists study the functions and structures of government and organizations, while specifically attending to the effects that these institutions have on their citizenry and in aggregate, society as a whole.

> Macroappraisal is thus a provenance-based approach to appraisal, where the social context of the record's creation and contemporary use (not its anticipated research use) establishes its relative value. (p. 128)

Cooks' articulation of macroappraisal was born in the aftermath of Canada's Deschênes Commission on Nazi War Criminals, which uncovered how a large number of immigration records were inadvertently destroyed by the Canadian National Archives. The controversy sensitized Cook to the abject failure of Schellenberg's concept of *informational value*, or the historical determination of record value. It led him to instead focus attention on present value, instead of future anticipated use, and to anchor that value on a determination of how citizens were impacted by an organization's activities. Cook drew heavily on German archival theorist Booms (1987), who argued against state controlled appraisal, under repressive Soviet-style communism, and advocated for a view of archival appraisal that measured the interests of society as a whole:

> In our view, a legitimate value standard or principle for the archival appraisal process can only be derived from this kind of contemporary valuation. Such a standard is inherent in history itself, for it is a standard of the past. It is not the product of speculation or ideological beliefs; it does not do violence to source material by applying value standards of the present which in the near future may already prove to be inadequate. If there is indeed anything or anyone qualified to lend legitimacy to archival appraisal, it is society itself, and the public opinions it

> expresses- assuming, of course, that these are allowed to develop freely. (p. 104)

Here we come full circle with the idea of archival appraisal as reflective of the goals of democratic societies as discussed earlier (Eastwood, 2002). Even with its focus on the state, its citizenry and society as a whole, macroappraisal foregrounds a particular governance structure that supports the needs of a democracy and even (in the case of Booms) notions of social opinion and the marketplace of ideas. While fully aware of the political role that appraisal takes in shaping memory, Cook understands appraisal as a tuning of the machinery of the state, to bring it in line with the interests and well being of its citizenry.

Booms' emphasis on the importance of public opinion is certainly understandable, and even laudable, when faced with state sanctioned repression, violence and media censorship. However an unquestioned reliance on a market for determining public opinion, without a critical engagement with the systems that generate those markets is problematic. Foucault's concept of governmentality is useful for unpacking and factoring these social and market forces, to see them not only at work in the functions of government, but also in the distributed material practices of individuals, and communities. Understanding how to measure popular opinion, or the impact of records on people, and some approaches that archives have taken to a form of appraisal that speaks directly to it political agency is where we turn next.

2.1.6. Appraisal Critique

In contrast to governance based approaches to archival appraisal are a group of divergent theories that speak directly to, and work to counter, the obscured, and often erased, influence and effects of power in archives. As we saw previously, structural approaches to appraisal often define appraisal implicitly by outlining principles of

original order and provenance that map to an institutional or societal context. In emphasizing the central role of provenance in structuring the archive, certain assumptions are made about what records get to count as archival, who creates them, and what resources are required to mobilize them as an archive. Another group of appraisal practices in archives speak to the archive as a site for record *creation*, not simply preservation. Archival records are part of a larger landscape of memory in which archival records circulate. These bottom up style approaches to appraisal archivists fully engage, and take political responsibility for, their role as shapers of the archival record, and the limits that their actions impose.

In a memorable, and oft quoted, address to the Society of American Archivists in 1974, Ham (1975) issued a wakeup call for archivists and their practice of appraisal:

> Our most important and intellectually demanding task as archivists is to make an informed selection of information that will provide the future with a representative record of human experience in our time. But why must we do it so badly? Is there any other field of information gathering that has such a broad mandate with a selection process so random, so fragmented, so uncoordinated, and even so often accidental? (p. 5)

Ham highlighted the lack of theorization around appraisal in order to make a plea for increased cooperation, empirical analysis, funding, and intellectual openness for archival appraisal. In many ways Ham was echoing a similar call by historian Howard Zinn, who addressed the same body a few years earlier saying that the archivist had a responsibility to document the lives of ordinary citizens and political movements, to hold power accountable through transparency, and to consciously work against the status quo in archives where

> … the existence, preservation, and availability of archives, documents, records in our society are very much determined by the distribution of

> wealth and power. That is, the most powerful, the richest elements in society have the greatest capacity to find documents, preserve them, and decide what is or is not available to the public. This means government, business, and the military are dominant.

Zinn's insights here continue to reverberate in archival studies especially when considering the process of appraisal. He foregrounds the hegemonic shape of archives, and challenges the archival community to not only hold the disciplinary form of the archive accountable, but also to move archival work outside of its governmental enclosure. In many ways this recognition of the role of power in shaping archives marks the birth of appraisal theory. It is in the moment when one can see the dominant mold of archives and its historical form that it becomes possible to talk about how that power has manifested in the records that have been collected and cared for in an archive (Jimerson, 2009). Foucault's exploration of the role of governmentality is instructive here because he lets us see how power works through the state, but also through disciplinary forms such as the family, the school, the factory, and the prison.

2.1.7. Strategy

Documentation Strategies is an appraisal technique developed by Helen Samuels as she worked to document university life at the Massachusetts Institute for Technology. For Samuels the archivist works as an analyst in a *distributed network* to study the social forces that generate records, rather than the records themselves. Samuels articulated four parts to developing a documentation strategy: 1) choosing a topic to be documented, 2) selecting the site for the strategy 3) examining the form and substance available documentation, 4) selecting and placing the documentation. This general strategy-for-developing-strategies, or meta-strategy, was distilled down to an appraisal technique known as Institutional Functional Analysis which bears some

resemblance to Cook's macroappraiasl, particularly in its attention the functions and societal impacts of institutions.

However, Samuels' crucial insight was that a *network* of analysts and repositories would be needed to address the challenge. A documentation strategy was not concerned with the decisions of a single archivist working within the hierarchical structure of a single institution. This networked approach mirrors Samuels' interest in emerging practices around the use of automated computer networks such as OCLC, RLG and at the distribution of machine readable cataloging data at Library of Congress (Samuels, 1986). A key aspect to Documentation Strategies is deciding where records are to be held, which may or may not be at the archivists own institution.

Significantly, Samuels also recognized that a documentation strategy may identify gaps in the documentary record, and could in fact suggest the *creation* of records where none currently exist.

> While archivists acknowledge the overabundance of information, they also recognize that modern communication patterns and records-keeping practices leave gaps in the documentary record. Documentation strategies, however, are ongoing activities and provide the opportunity to intervene in the records creation process and assure the creation and retention of required information. (Samuels, 1986, pp. p121–122)

Seeing archival appraisal as an intervention and as a potentially creative act was a surprisingly radical move, especially given the vast number of records already in need of preservation, and the anxiety about preserving them.

2.1.8. Decentralization

This shift in attention outside the walls of the physical archive suggested by Documentation Strategies was in fact part of a longer movement in archival studies that often gets labeled *postcustodial*. The term postcustodial was first used by Ham (1984) in drawing attention to the ways in which information technology and automation were dramatically transforming the landscape of record production, and the concomitant need for the archival studies field to invest in researching these new forms of record production, in order to adequately preserve and provide access to them. A decade later Cook (1993) synthesized a significant body of work by Taylor (1988), Bearman (1989), Hedstrom (1991) under the rubric of *postcustodial*, which he aligned with post-modern theories of archives. For these postcustodialists a transformation of archival theory, and especially appraisal, was absolutely necessary because of the proliferation of electronic media, which resist the idea of a singular documentary artifact.

> In this fluid electronic environment, the idea of a record physically belonging in one place or even in one system is crumbling before new conceptual paradigms, where "creatorship" is a more fluid process of manipulating information from many sources in a myriad of ways, or applications, rather than something leading to a static, fixed, physical product. For information professionals, this signals that the custodial era is giving way to a postcustodial one, where the curatorship of physical objects will define our professions much less than will an understanding of the conceptual or virtual interrelationships between creating structures, their animating functions, programmes, and activities, the information systems, and the resulting records. (Cook, 1993, p. 424)

Postcustodial thinkers emphasize that appraisal is not simply a valuation of records

during record acquisition, but is inherent in the design and construction of information systems: "archivists need to reexamine how information systems support organizational functions and relate to organizational structure within specific organizations and in a broader documentary context" (Hedstrom, 1991, p. 344). Postcustodial archival theory crystallized in the Records Continuum model, which dismantles the idea of a linear life cycle for records in which the archive is the place where records come to rest. The Records Continuum instead stresses how archival records are part of multiple, recursive, processual flows, as they cycle between creation, capture, organization and pluralization (McKemmish et al., 2010). In the Records Continuum appraisal is less concerned with ascertaining the value of records as it is with the design and implementation of *systems* that generate the records.

2.1.9. Outside the Archive

The need for appraisal to encompass the record creation process is also reflected in community archives approaches. Community archives extend and build upon postcustodial archival practice by situating the work of appraisal out in the world, in particular social contexts where records are created and actively used (Flinn et al., 2009). This movement allows communities to retain custody of their records, which affords more autonomy during the appraisal process. The decisions of what to archive are not being conducted solely by archivists, but also by members of a particular community of people, who are ultimately using the records. Rather than simply treating custody as something to decenter or transcend, because of the material configurations of computer networks and information technology, custody is directly linked to access and the use of records (Bastian, 2001). Community archives allow for records to take on new historical dimensions, interpretations and use for identity formation (Punzalan, 2009) and collective memory (Brothman, 2001 ; Jacobsen et al., 2013).

It is significant that community archives approaches often develop in response to the systematic erasure, marginalization, or disenfranchisement of particular groups, that is performed by traditional, institutional forms of archives (Flinn, 2007). In this way community archives speak to the archive as a site and instrument of power (Jimerson, 2013), for social justice (Punzalan & Caswell, 2016) and even activism (Cifor et al., 2018). In the context of social justice, archival appraisal fully admits, and even celebrates, its active political role in shaping and reshaping collective memory and identity (Harris, 2002). For some, information technologies, such as the web and social media, offer new opportunities for participation, which broaden the set of actors who can perform appraisal, and thus the diversity of records (Gilliland & McKemmish, 2014; Huvila, 2008).

More recently there has been a move to position archival appraisal in terms of a feminist ethics of care, which de-emphasizes a rights-oriented, legalistic approach to social justice, which can inadvertently work to reinscribe the very same oppressive systems they are working to dismantle (Caswell & Cifor, 2016). The ethics of care approach expresses appraisal and other archival activities in terms of a network of mutual responsibility. These relationships become a conduit for appraisal as a measure of affective value, for deepening engagement with communities and furthering social justice (Cifor, 2016). We will turn to a deeper discussion of the ethics of care and its role in understanding web archiving as a set of repair practices shortly.

2.1.10. Appraising Appraisal

Admittedly, this cursory overview has glossed significant aspects of archival appraisal, while also failing to mention others. However my goal here was not to provide an exhaustive description of the field, but to describe a constellation of divergent and even opposed archival appraisal theories, in order to situate them with Fou-

cault's notion of governmentality. Governmentality helps us to examine appraisal theories as a set of knowledge practices, that orient around organizational structures as well as to individual agency, and reproduce themselves at the levels of societies, states, communities, and even individuals. Appraisal is implicit in the duty to record the activities of the state or administrative body; it works along the grain of organizational hierarchies to document what is deemed most important; and it operates in the service of documenting society, and the interests of democracy and its citizenry. But in a counter movement appraisal also works to decenter the archive as institution, and to locate appraisal practices in systems of record creation and the design of information systems. Appraisal responds to the juridical form of the archive: to create records where none exist, and to reinterpret existing records as part of recursive process that dismantles linear conceptions of the record lifecycle. Appraisal is a tool for social justice that speaks directly to power formations, in order to address oppression and marginalization, while imagining new frames for collective memory. Appraisal even attempts to reconcile the dialectical forces of structure and agency by centering the ethics of care and practice, and moving outside of western human rights based frameworks.

2.2. Appraisal and Web Archives

And so we return to the question that we started with: how are we deciding what to archive from the web? Or, how does the constellation of archival praxis around appraisal that we have just discussed meet up with web archiving practices? You will notice that so far there has been very little discussion of the actual materials that are the subject of appraisal. It's almost as if our theories of appraisal are thought to function independently of the material being appraised; that in principle (if not in fact) appraisal can be applied as needed to all forms of media. Indeed, this abstractedness is part of appraisal's claim to theory in the first place, and forming part of an

archival *science*.

However the birth of appraisal, or the awareness that records *must* be selected, in some fashion, by the archivist, occurs in a particular historical moment of profound material transformation, as the technologies of record production completely outstrip the archivists ability to process their outputs. This *overcoming* of the archivist, and the concomitant appraisal response, are not simply the result of an increased volume of records, or information overload (Ham, 1981). The concept of archival appraisal arrives as the centuries old archival technologies of paper, print and their containers are giving way to a proliferation of electronic media formations, which challenged and continue to challenge the archival imagination (Taylor, 1988). Cook (1994) suggests that appraisal is perhaps more of a sensibility than an abstract theory, for the way it offers an approach to practice, and a consolation amidst this transformation:

> Most important, for the first time, we are not producing, managing, and saving physical things or artifacts, but rather trying to understand and preserve logical and virtual patterns that give electronic information its structure, content, and context, and thus its meaning as a "record" or as evidence of acts and transactions.

Here Cook casually deploys the idea of "virtual" records and the effects that they are having on archival practice. This virtuality derives at least in part from how computer technology collapses the media forms of word, picture, sound, video into a seemingly uniform binary representation–the so called ones-and-zeroes of digitality. However electronic records, especially born-digital-records, of which the web is a prime example, have actual material dimensions, and condense as the result of processual flows that involve platforms, infrastructures, networks, heuristics and algorithms. The virtuality of electronic records can be understood as what Kirschen-

baum (2008) calls a "screen essentialism" or medial ideology that effaces the instrumentation that creates, and must constantly recreate, the experience of an electronic document. At the same time, archives of the web are particularly prone to a technological solutionism that treats preservation as a purely technical problem, where we simply need more efficient and less error prone storage, more comprehensive crawling strategies, or improved network protocols in order to "archive it all".

Despite efforts to archive the entire web (Kahle, 2007), the idea of a complete archive of the web remains both economically infeasible (Rosenthal, 2012), and theoretically intractable (Masanès, 2006c). Features of the web's Hypertext Transfer Protocol (HTTP), such as code-on-demand (Fielding, 2000), content caching (Fielding et al., 2014) and personalization (Barth, 2011), have transformed what was originally conceived of as a document oriented hypertext system into a complex multimedia information infrastructure that delivers content based on who you are, when you ask, and what software you use (Berners-Lee & Fischetti, 2000). As a result, the very notion of a singular documentary artifact, which has been under significant strain since the introduction of photography (Benjamin, 1999) and electronic records (Bearman, 1989), is now being pushed to its theoretical and conceptual breaking point. We rarely try to reason about *all* archival records that exist in the world as a singular assemblage. So why do we do this so naturally with the network of networks that we call the Internet, or the massively distributed text that is the web? Dissolving the seeming virtuality of web archives, and understanding both the web and web archives as social, historical and technical artifacts is essential for being able to talk about how web content is being selected for preservation (and access) in an archive. Now that we are 30 years into the web's existence we are in a position to turn from the innovation-centric approaches to web archives, to the critical investigation of established practices.

2.2.1. Science and Technology Studies

In the early days of the web Hedstrom (1991) proposed a broad research framework for the study of *electronic records*, or what some today might call computational archival systems. This framework was based on a *sociotechnical* understanding of archives that recognized the need for archival studies research to move beyond technological determinist accounts, where society is shaped by technology, as well as overly socially determinist accounts, where technology is shaped purely by social and political interests. In *sociotechnical* accounts both technology and society must be understood in terms of a complex interplay between people, materials and processes, where the technical and the social mutually shape, or coproduce, each other. This line of inquiry is especially relevant today,and builds upon the rich empirical foundations of Science and Technology Studies (STS) (Felt, 2017; Latour, 1987 ; Pinch, 1987).

Hedstrom's framework for research into electronic records included 1) the relationship between automation, electronic records and organizational change; 2) new material forms of electronic records and their remediation of previous forms; 3) the design of new information systems with archival properties; 4) evolving markets for information technology and their impacts on archival records; and 5) the impact of electronic records on accepted norms and approaches to archival preservation (e.g. provenance and appraisal). Hedstrom stressed how attending to the social, historical and material dimensions of information technology, was the cornerstone of archival studies:

> The introduction of new forms of material and the simultaneous transformation of traditional forms into something new raises a series of questions about the relationship between forms of material and archival practice. When should new forms of material be managed differently

> from more traditional forms of documentation? Are there any archival principles that apply to all new electronic record types? What characteristics does an electronic memo share with a memo on paper? What does it have in common with other machine-readable records?

Hedstrom's framework holds up particularly well today, as many of these five areas have developed into full fledged fields of study. But this development, at least for the study of web archives, has mostly happened either in the domain of computer science where the concept of appraisal is infrequently used, or in the field of archival studies where the web is considered under the rubric of electronic records. Before discussing how STS can be applied to the study of appraisal in web archives I will briefly characterize these literatures and their perspective on web archives.

2.2.2. Digital Libraries

The computer science literature reflects a decade long interest in *digital libraries*, particularly the research output of Joint Conference on Digital Libraries (JCDL), that for the last 20 years has served as a research forum for all manner of investigations into novel methods for collecting, storing, indexing, accessing and preserving content collected from the web. When it comes to appraisal, digital library research on the subject of web archives has focused in large part on the problems of *harvesting* or *crawling* the web. Practices for crawling the web for the purposes of indexing and maintenance are nearly as old as the web itself (Fielding, 1994). However the actual dimensions of the web are still difficult to determine (Dobra & Fienberg, 2004), and the "deep web" presents challenges because of the way large regions of the web are hidden behind human driven query interfaces (Madhavan et al., 2008).

Substantial work has been done to measure the *archivability* of web content (Banos & Manolopoulos, 2015 ; Samar et al., 2014) and to determine the age of web re-

sources (SalahEldeen & Nelson, 2013), which are important factors to consider when deciding what to archive. Another significant consideration is the expected disappearance of a resource from the web. Numerous studies have investigated the transience of web resources, so called link decay, link rot, or reference rot, in order to characterize the likelihood of certain types of web content disappearing. Measures of link rot can be used to mobilize efforts at increased web archiving generally, as well as specific areas such as social media (Salaheldeen & Nelson, 2013) , science (Hennessey & Ge, 2013), institutional repositories (Sanderson et al., 2011), and jurisprudence (Zittrain et al., 2014).

The problem of knowing what to collect from the web has also been treated in the digital library research community as a *focused crawling* problem. In focused crawling the goal is to collect content about particular topics (Risse et al., 2012), events (Klein et al., 2018; Yang et al., 2012), or to collect content that has a particular characteristic such as popularity (Page et al., 1999), importance Baeza-Yates et al. (2005)] or social engagement (Gossen et al., 2015 ; Milligan et al., 2016; Nwala et al., 2018). Generally speaking these approaches take the focus to be a topic, event, person, organization that can be qualified by the types of media (documents, audio, video). These criteria are taken as a priori, or a given, that is decided out of band, before the task of doing the crawling is undertaken. Similarly there is also a vast research literature on relevancy ranking in web search results that is dependent first on having already amassed a collection of web documents to index. However how that corpus of documents has been assembled is usually taken as a given. Despite the lack of the use of the term appraisal, this vein of digital library research speaks to Hedstrom's call for archival studies to directly engage in the design of record keeping technologies for the web.

More recently Lee has highlighted the lack of research into what he calls "computer

assisted appraisal" where digital forensics tools, natural language processing (NLP), machine learning (ML) are applied to appraisal activities (Lee, 2018). Lee observes that these tools have typically been used in digital library and archives as a means of providing access points, where personal names and topics are algorithmically generated from archival records in order to assist researchers in using records. However Lee points out that these tools can potentially be used to assist in the selection of records for archives. For example, time metadata can be extracted from records using digital forensics tools so that content can be selected for an archives based on a timeline. In addition, previous appraisal decisions can be used as training data for supervised machine learning algorithms that can automatically select content going forward as new records are acquired. Coincidentally Lee cites research presented here in Chapter 5 as evidence of appraisal in web archives being well studied. While a case can be made for this when it comes to the application of information technology in the area of archival appraisal, there is still much work to be done to study the sociotechnical factors at play in "computer assisted appraisal", which do not privilege algorithmic approaches.

2.2.3. Electronic Records

Within the field of archival studies there are numerous accounts of how appraisal applies to electronic records, but less so with regard to the web specifically. The term *electronic records* developed alongside computer automation to refer to the databases, electronic mail archives, and other types of data that accrue as *files* on a computer file system. These computer filesystems used the metaphors of paper based documents, files and containers for storing them, in order to make familiar a completely new information infrastructure. Designating these computer files as electronic records highlights how computer data is generated as part of the functioning of organizations, and provides evidence of that operation, in which authenticity,

reliability and fixity are typically paramount (Cobb et al., 2005 ; Duranti, 2010 ; Harvey & Thompson, 2010). Discussion of electronic records also happens under the rubric of digital preservation (Rothenberg, 1999) or digital curation (Yakel, 2007). In addition to the preservation of computer files and file systems researchers have worked on means of format identification (Hitchcock et al., 2007) and pursued the use of software emulation to preserve and describe entire software systems (McDonough et al., 2010; Welch et al., 2012). C. A. Lee & Tibbo (2011) suggest that the term *digital curation* reflects a postcustodial response to digital preservation, that foregrounds the site of record production as outside the walls of the archive, instead of focusing entirely on the authenticity and fixity of data, and conventional ideas of the preservation lifecycle.

Electronic records, digital preservation and digital curation are largely predicated on prior custody of data, or the ready access to the computer systems (software and hardware) that the data has been produced on. These approaches tend to take as a given that the data is in fact available, and generally do not speak directly about the process of archival appraisal, or the social and technical means by which data is selected for preservation in an archive. As such the web presents a conundrum of sorts, where content *appears* openly available, and is often regarded as *public*. But even as they seem ready-to-hand, web documents also exist at a distance, sometimes in multiple locations, and are retrieved across blurred geopolitical boundaries, which the infrastructure of the Internet and the architecture of the web (when working) makes instantly available in the web browser.

Many accounts of the appraisal of web content take a case study approach where web content is collected to document a particular event or category of content. For example Masanès (2006c) examines the identification and crawling of websites relating to the French presidential and parliamentary elections in 2002 by the Bibliothèque

nationale de France. In another prominent study Schneider et al. (2003) detail approaches to creating a collection of web content at the Library of Congress related to the September 11 terrorist attacks. Changes to copyright law enabled national libraries to evaluate both broad and focused approaches to archiving entire top-level-domains (e.g. .fr or .pt) (Lasfrgues et al., 2008, and @Gomes:2006). The need for tools that allow *seed lists* (lists of URLs to archive) to be developed was articulated by Pearce-Moses & Kaczmarek (2005). Duncan & Blumenthal (2016) describe how networks, such as professional consortia, are useful resources for curating seed lists. In addition there has been discussion of the use of social media and their Application Programming Interfaces (APIs) as a means for discovering web content for archiving events such the Arab Spring (Arnold & Sampson, 2014) and the protests in Ferguson and the #BlackLivesMatter social movement (Rollason-Cass & Reed, 2015). Web archives are also considered as a site for critical engagement with issues of social justice (Aronson, 2017).

In one of the more conceptual models of appraisal in web archives Masanès (2006a) analyzed the process of *selection*, which is broken down into three different phases: preparation, discovery and filtering. Preparation involves establishing a focus for the collection, as well as selecting the technology for performing the collection. Discovery involves the act of collection itself, which leverages both the structure of the web (endogenous) as well as external resources such as search engines and link hubs (heterogeneous). Filtering is a process by which collected material is selected for an archive based on quality, subject, genre and publisher. Masanès discusses the degree to which automation can be used in these three phases, but unlike the previously mentioned case studies the discussion is prospective and largely unanchored from actual practice. For example, whether and how archivists are involved in the filtering of collected web content is not explored.

The architecture of the web presents blurry geographies, where it is not always clear where one website ends and another begins. The idea of a singular document, which is central to information science (Buckland, 1991), is put under strain by the web's use of hypermedia, which allows documents to be dynamically composed from heterogeneous sources. Practices for appraising the web diverges significantly from electronic records in that it necessarily involves collecting or assembling content from the Internet. Appraising web content also requires the means and mechanisms for reassembly, or replay, of the content later (Andersen, 2013) in order to assess what has been collected. This recursive assembly and reassembly of content is tightly bound up with the appraisal process itself, and is achieved with specialized tools (Mohr et al., 2004), that have particular design assumptions, goals and affordances built into them.

Zooming out from the technical specifics, appraisal in web archives also can be seen as a form of virtual reunification (Punzalan, 2014) where who the various stakeholders are, their conceptions of process and product, and goals in creating the archive are made manifest. However in the case of web archives this reunification was always already virtual, at least in the sense that they were very often born digital, as part of the accrual of data and documents in the network. Unpacking the very real social and material processes and practices that underlie the virtuality presented by the web archive is a key component to understanding how archival appraisal on the web works.

2.3. Appraisal Practice

So far I have outlined how ideas of archival appraisal have developed over time, with the goal of showing how these conceptions are used (and not used) in the research literature of web archives. In examining the research literature of appraisal I have drawn on Foucault's idea of *governmentality* as a way of understanding how

a wide set of archival appraisal practices reflect, repeat and respond to concerns of governance. However, for Foucault governmentality is about much more than simply the affairs of government and state: it also includes the practices of individuals, families and groups of all kinds:

> This word [government] must be allowed the very broad meaning it had in the sixteenth century. "Government" did not refer only to political structures or to the management of states; rather, it designated the way in which the conduct of individuals or of groups might be directed: the government of children, of souls, of communities, of the sick. It did not only cover the legitimately constituted forms of political or economic subjection but also modes of action, more or less considered or calculated, which were destined to act upon the possibilities of action of other people. To govern, in this sense, is to structure the possible field of action for others. (Foucault, 1982, p. 790)

Foucault developed the concept of governmentality by examining how institutions like hospitals, schools, prisons, barracks, and even archives (Foucault, 1986) work to discipline, and delimit the field of action. Foucault's examination of specific practices and their relations is considered by many to be part of a general turn to practice in social theory. Postill (2010) identifies Foucault as part of a first wave of practice theorists that also includes Ludwig Wittgenstein, Pierre Bourdieu, Anthony Giddens, and Michel de Certeau. He characterizes these thinkers as working to:

> … liberate agency – the human ability to act upon and change the world – from the constrictions of structuralist and systemic models while avoiding the trap of methodological individualism. These theorists regarded the human body as the nexus of people's practical engagements with the world. (p. 7)

Practice theorists are interested in resolving the dialectic between individual agency and the social and material structures that constrain and reproduce it, through a material examination of the processes and routines that center on our physical bodies and experience. Giddens (1984) uses the idea of *structuration* to propose that agency and structure are mutually constitutive entities with equal status, each of which recursively reproduces the other. The idea of structuration has already seen some purchase in the field of archival studies where it has been used for understanding appraisal (Brown, 1991), descriptive practices (Yakel, 2003), the space/time dimensions of records (Upward, 1996 , 1997), and the use of collaborative documentation networks.

Similarly Bourdieu (1977)'s idea of *habitus* identifies the site of our lived experience which is shaped by social structures, which generate individual action. Schwartz (1995) mobilizes Bourdieu when analyzing how archival photographs function as documents that both produce, and are the product of, social rules. Gracy (2007) has also drawn on Bourdieu's notion of field, habitus and social capital in researching how commercial and nonprofit film archives operate. In some significant recent work Ivanov (2017) conceptualizes archival theory in terms of practice theory, and uses both as a framework for examining record keeping practices in large news organizations.

At first it might seem strange to consider the physical body and its practices in relation to something as seemingly immaterial as the web. But the web and the underlying Internet are physical infrastructures that constantly generate records as we point and click in our browsers, swipe the screens of our smartphones, as we are quietly surveilled by the Internet of Things (Acker, 2015). Archives of web content sediment on servers as we document our lives using record *making* technologies such as Facebook Live, which aren't always record *keeping* technologies (Sheffield,

2018). Given this orientation towards practice and the general theoretical umbrella provide by Science and Technology Studies I will conclude by suggesting several promising areas for for future research into appraisal and web archives.

2.3.1. Ethnography of Infrastructure

The ethnographic study of infrastructures in terms of the human practices that play out in their construction and use is well developed in information studies (Bowker, 2005; Star, 1999 ; Suchman, 1985). For example Edwards et al. (2011) has examined how metadata practices work to shape knowledge production in the sciences. The mutually constitutive roles of policy, practice and design can shape infrastructures and define the controversies that surround them (Steven J Jackson et al., 2014). Social media applications like blogs have been studied for their role in shaping scholarly communication infrastructure (Burton, 2015). The *convivial decay* of space science infrastructure has been investigated by Cohn (2016), who found that notions of repair, aging and multiple intersecting life stories of people & machines played a significant role in how infrastructure is designed, deployed, used and then dismantled. Karasti & Blomberg (2018) propose a methodological framework anchored in STS and anthropology for analyzing infrastructures over long time scales, that we see evidenced in the work of Ribes & Finholt (2009) on the *long now* of infrastructures.

The ethnographic study of web archiving infrastructure in particular has also begun, but more work remains. Ogden et al. (2017) provides a rare glimpse at how labor and infrastructure intersect in the practices of web archivists at the Internet Archive. Similarly Kosnik (2016) has performed a detailed ethnography of archival practices in fan fiction communities on the web. This recent work connecting web infrastructure, archives, and social practices suggests further ethnographic investigations of how web archives both shape and are shaped by ideas of appraisal, or what is deemed valuable in the web.

2.3.2. Repair and Maintenance

The turn to practice in Human and Computer Interaction (HCI) has been noted by Kuutti & Bannon (2014), who suggests that HCI's focus on user needs, or macro-level organizational systems are no longer sufficient for understanding the complex formations of users and systems we see today in digital environments. The study of computer systems in terms of practice offers a way of collapsing these scales. Kuutti points out that the turn to practice in HCI research takes shape around issues of performativity of social practices, materiality of human bodies and artifacts, agency, and knowledge generation.

The role of repair as a site for design, in contrast to more conventional and celebratory notions of innovation, is detailed in ethnographic work by Steven J. Jackson & Kang (2014), DiSalvo et al. (2010), Rosner & Ames (2014) and Maestri & Wakkary (2011). *Broken world thinking* (Steven J. Jackson, 2014) specifically recognizes that design happens not only in the experimental setting of HCI research, but in our everyday lives as we learn to adapt and improvise solutions to conceptual and infrastructural breakdowns. Repair and breakdown are seen as two sides of the same coin of maintenance. Russell & Vinsel (2018) situate maintenance studies in terms of the history of technology, where the investigation of how standards are developed can help make sociotechnical assemblages legible (Lampland & Star, 2009; Russell, 2014).

With the maintenance perspective in mind it is possible to broaden the scope of what appraisal means in web archiving systems to include not only the design and use of web archiving systems, but also the material practices and labor that sustain the web, as software is upgraded, vendor contracts are abandoned and content is ported from one system to another. Tracking these processes of data migration as appraisal decisions in web archives is a viable and under explored avenue for archival studies

research, particularly in connection with postcustodial theories of the archive.

Hedstrom (2002) suggests that the archivist's appraisal decisions are an essential part of the archive's *interface* with the past, found both on the screen, and embodied in the archivist. In some recent work Maemura et al. (2018) has begun to explore how archival decisions manifest in the provenance of web archival systems, in terms of the documentation they provide and the sociotechnical means by which that documentation is generated and conveyed. Further surfacing the repair work of archivists, and the web as a site for continual repair and maintenance is a key area for future research into archival appraisal on the web.

2.3.3. Software Studies

The field of software studies provides a humanistic method for reading software and digital media systems as artifacts, with particular material, social and historical dimensions. For example, Manovich (1999) has studied the database as a narrative form whose ontology structures the cultural artifacts that it helps generate. Kelty (2008)'s pioneering ethnographic work has helped us understand open source software development communities as a *recursive public* that uses the infrastructure of the web and the Internet to develop and express what the Internet is, or can be.

Another class of research that fits (perhaps a bit less comfortably) under the rubric of software studies is work that explores the materiality of digital media in terms of inscription (Kittler, 1999), memory (Chun, 2011), digital storage and transmission (Kirschenbaum, 2008), protocols (Galloway, 2004), physical networks (Starosielski, 2015), documents (Gitelman, 2014; Levy, 2001), data representation (Dourish, 2017) and even archives (Ernst, 2013). In this varied literature there is a consistent engagement with how data systems are anything but neutral (Bowker, 2005 ; Gitelman, 2013 ; Walford, 2017), and are configured by the material that they are

constructed from, and by the design decisions of their creators and maintainers.

One relevant area for understanding appraisal in web archives is the growing fields of platform and algorithm studies, which are closely aligned, but operate from different levels of abstraction and granularity. Platform and algorithm studies examine the social, political and cultural contingencies that the technologies provided by organizations like Google, Facebook, Twitter, Wikipedia coproduce with their publics and users. At the platform scale these contingencies manifest in established policies such as moderation rules, community guidelines, and terms of service documents, which are translated into actual practices, and in turn generate political economies (Gillespie, 2010 , 2018). Similarly algorithm studies, despite its usual association with computer science, provides a humanistic lens for studying the development, use and impact of computation in particular settings, enabled by specific practices, in order to achieve explicit or tacit ends (Gillespie & Seaver, 2015 ; Seaver, 2017). As such platform and algorithm studies fit within the scope of critical data studies (boyd & Crawford, 2012), which shifts the focus from policy and code to flows and accumulations of data, and their politics (Bratton, 2016; Zuboff, 2015).

As noted by Dourish (2017), it's awkward, and perhaps a bit misleading, to lump all these theoretical concerns into the category of *software studies*. However doing so, is a convenience that allow us to talk about the possible avenues of future research for web archives, especially with respect to appraisal. At the risk of introducing yet another category, Kitchin & Lauriault (2014) draws on STS terminology to mobilize the idea of *data assemblages*, which are heterogeneous, sociotechnical constructs that resemble Foucault's *dispotifs* that act as bundles of "discourses, institutions, architectural forms, regulatory decisions, laws, administrative measures, scientific statements, philosophical, moral and philanthropic propositions" (quoted in Kitchin & Lauriault, 2014). Kitchin suggests several methods for studying the complex

phenomena that are data assemblages:

1. examining code artifacts, their data inputs and outputs as time bound processes
2. reflecting on the writing & design of code, heuristics and algorithms
3. reverse engineering algorithms, to intuit processes and architectures that are hidden from view
4. conduct ethnographies of design teams to ascertain the contingent, relational, and contextual way software is produced
5. widening the ethnographic lens to consider institutional and organizational forces at work
6. widening the ethnographic lens yet again to consider the work that the algorithmic systems do in the world, either intended or unintended, and their social and historical significance

In some significant recent work Ben-David & Amram (2018) used techniques from the study of *black box* algorithms (algorithms whose inner workings are secret, or so complex that they are unknowable) (Diakopoulos, 2014) to consider the epistemic role of web archives as fact or evidence producing systems. Ben-David specifically looks at the representation of the the North Korean top-level domain (.kp) in the Internet Archive's Wayback Machine, using traces of provenance information provided by the Wayback Machine's interface, in combination with the historical contingencies of DNS leak that occurred in 2016. Ben-David finds that the Wayback Machine's processes for acquiring data from the web are found to be iterative, top-down, bottom-up, and that they extend laterally outside of the Internet Archive organizational walls.

Similarly Schafer et al. (2016) employs techniques from STS to unpack the sociotechnical black-boxes of web archive infrastructure to stress the importance of

web archives for Internet governance. Both pieces highlight how the quiet labor of web archives, performed by archivists in collaboration with, and sometimes in spite of, machines, are of historical and political significance. More work remains for understanding the technological, social and historical contexts in how these software systems, standards and organizations have developed–particularly with regards to our understanding of what a web archive is, what it is not, and how the difference is negotiated. The time is ripe for these analyses since we have 30 years of web, and 20 years of web archiving practice to study.

2.3.4. Data Justice

And finally, a significant strand of work in the turn to practice has drawn attention to how the design of information systems can respond to the needs of social justice. The imperative for technology to be designed by those who are supposed to ultimately use it was initially developed in the 1960s in Scandinavia, and gave rise to the field of participatory or cooperative design, which has garnered sustained interest from the HCI and Computer Supported Cooperative Work research communities (Ehn, 1988 ; Kensing & Blomberg, 1998; Star & Ruhleder, 1996 ; Suchman, 1985). Related concepts such as Communities of Practice (Lave & Wenger, 1991) provide a rich methodological framework for understanding how social practices involving technology constitute the way we learn and work together in sustainable ways. Furthermore, the role of information technology infrastructures in measuring and making legible environmental change (Edwards, 2010 ; Steven J Jackson & Buyuktur, 2014) while also negatively impacting local and global environments (Cubitt, 2016) is an increasingly important area of concern. Indeed, the general thesis that there is no escaping the politics that are embedded in our technologies is a theme that is returned to again and again (Winner, 1980). More recent recent work has explored how data infrastructures are both the product of, and can give rise to

capacity building for mobilizing change (Meng & DiSalvo, 2018; Tufekci, 2017). This work uses an ethnographic approach to consider how data practices are part of a constellation of other social and political practices.

This research literature bundled under the theme of social justice in information technology intersects broadly with archival studies in significant ways. As noted by Punzalan & Caswell (2016), archives and memory studies have been perennially interested in the role of social justice. These researchers have worked to dispel the notion that archives are neutral in their representation of communities or society as a whole, and that they necessarily embody particular social and political values. Specifically the goals of community archives (Flinn, 2011) and participation in archival processes of appraisal and description have been marshaled from a variety of angles (Gilliland & McKemmish, 2014 ; Huvila, 2015) that mirror the goals of participatory design.

In terms of addressing web archives, and specifically the needs of archival appraisal, there has been some emerging work, but much remains. Goldman (2018) has begun studying how digital preservation frameworks such as the Open Archives Information System (OAIS), and principles such as Lots Of Copies Keeps Stuff Safe (LOCKSS), can introduce overheads that negatively impact the environment, and if widely practiced, could undermine the long-term sustainability of data archives. Christen et al. (2017) has explored how values expressed in open access licenses and metadata standards can be at cross purposes with the ethical curation and access to cultural heritage materials. These risks can be mitigated by engagement with local communities and community archives practices. Furthermore, in the wake of the 2016 US Presidential Election the Environmental Data Governance Initiative worked to preserve at risk data sets produced by the Environmental Protection Agency and other US federal agencies. This work was articulated as a web archiving

project, and was performed by volunteer archivists, librarians, technologists, scientists and concerned citizens (Schlanger, 2017). Participants have reflected on how this *data justice* work fits into an evidence based accountability framework (Dillon et al., 2017). But understanding how these and other emerging data justice projects (Taylor, 2017) take place both in and through the infrastructure of the web, and express implicit or explicit appraisal decision is work that remains.

2.4. Appraisal

My review has drawn from the archival studies, digital library and sociotechnical research literatures in order to highlight the gap in our understanding of web archival appraisal as a sociotechnical practice. The reasons for this gap are the result of a general lack of empirical studies of how materials are selected for web archives, where the focus isn't so much on the social practices or the technical practices, but on the diverse set of interactions that occur when they are blended. Foucault's theory of governmentality, or the conduct of conduct, bridges this conceptual terrain, and offers a way of studying appraisal in web archives that aligns with current notions of critical algorithm studies and critical data studies. Providing an empirical study of how web archives are constructed is the principal goal of this dissertation. It is in the analysis of material working practices in web archives where appraisal values can be identified and theorized.

3. Methodology

In the previous chapter I outlined the various ways that archivists have historically conceived of *appraisal*, including its more recent expression in the domains of digital curation and digital preservation. This review was a necessary first step in beginning to pose my specific research questions which concern how appraisal is being enacted in web archives, and how web archives are conceived. What emerged from this review is that it is useful to understand archival appraisal not only as a product of governance (a concern of organizations, institutions and the state), but also through what Foucault calls governmentality, or the microphysics of power. While it is increasingly accepted to conceive of archives as a technology of power (Jimerson, 2009; Zinn, 1977), the exercise of archival agency is found not only in the top-down operations of governments and institutions, but also in every day practices at multiple, dispersed sites spread throughout society: in workplaces, families, communities and *collectives* of all kinds, and even by individuals (McKemmish & others, 1996). The concept of governmentality allows us to investigate these sites of archival appraisal as part of a continuum. Governmentality shifts our focus to the *practices* of appraisal and how they can enact power relations and resistance as well as social justice and collective memory (Jacobsen et al., 2013; Punzalan & Caswell, 2016).

An additional related theme that the previous chapter introduced is that the question of how appraisal operates in *web* archives requires an analysis that accounts for the role that technology plays in these practices. Of course, technology has always been an active agent in record keeping. This in itself is nothing new. And yet we are only a few decades into a continuing and profound shift to electronic records, in which centuries old archival technologies of paper have given way to a digital regime of databases, data processing, and computer networks (Bearman, 1989). As

Hedstrom (1991) outlines in her *Framework for Research on Electronic Records*, the field of Science and Technology Studies (STS) provides a useful historical and theoretical framework for understanding archives as sociotechnical systems, without privileging either social or technical explanations. What gets deemed *archival*, and the very meaning of *preservation* and *access* are forged in the design and use of information processing systems, and attendant standardization practices. Similarly, the practices of *appraisal* in the context of the web, and even the definition of what constitutes a *web archive*, are the result of sociotechnical processes in which our design of algorithms, data formats and interfaces both produce and are produced by web technologies. STS provides both a theoretical and methodological framework for researching the ontological dimensions of appraisal.

So, my research project theorizes appraisal in web archives as a sociotechnical practice, where these practices can be understood using the framework of governmentality. This chapter outlines a methodology for investigating the question of how archivists decide what to archive from the web, and what significance these findings have for larger questions about memory, technology and the web. The processes of how we choose to remember with the web are critical for ascertaining what our web archives *mean* (Ketelaar, 2001 ; Maemura et al., 2018). However, it is important to remember that us humans are not the only readers of the records in web archives. The "meaning" of web archives is increasingly found in human-computer assemblages that are designed to "learn", or establish statistical patterns in archival data, so that very real decisions can be made in the world (boyd & Crawford, 2012 ; Mackenzie, 2017). Web archives are now assembled by machines, in order to be read by machines, for very human purposes. The philosophical question of whether machines actually "understand" such records or not is increasingly irrelevant, but the stakes for our understanding how records come to be in our web archives could not be higher.

3.1. Algorithms as Culture

As sociotechnical assemblages, web archives are complex sites where humans and computers collaborate to select web content for an archive. Indeed, on close inspection it can often be difficult to untangle these relations and clearly demarcate where one begins and the other ends. Web archives are infrastructures of software and hardware that have been crafted by archivists and technologists over the past two decades to do something we've come to call "web archiving". Archivists and other types of users interact with these systems to collect particular regions of the web, and these systems continue to change to help the further achievement of those goals. Finding techniques and methods for bringing these blurry boundaries and knotted relations of web archiving into focus is the task at hand.

Instead of being purely virtual, disembodied and abstracted, *algorithmic* processes are the result of distinct social and material practices, in very specific and highly contingent settings (Geiger, 2014). I have introduced the term *algorithm* here because, as we will see, the appraisal of web content is increasingly automated by crawling procedures or heuristics. I also want to use the methods of *Critical Algorithm Studies* to help answer the question of how appraisal operates in web archives. Critical Algorithm Studies centers the study of algorithms as material artifacts that participate in particular social settings.

Algorithms have traditionally been considered the sole domain of the computer scientist. For example here is a conventional definition of the term algorithm found in a popular undergraduate computer science textbook:

> … any well-defined computational procedure that takes some value, or set of values, as *input* and produces some value, or set of values, as *output*. An algorithm is thus a sequence of computational steps that transform the *input* into the *output*. We can also view an algorithm as a

> tool for solving a well specified computational problem. The statement of the problem specifies in general terms the desired input/output relationship. The algorithm describes a specific computational procedure for achieving that input/output relationship. (Cormen et al., 2009, p. 5)

Notice how this definition works to scope the concept of the algorithm to a particular setting: where an algorithm is a "tool", that helps solve a "problem" in terms of set of "inputs" and "outputs". The "problem" is computationally constrained, and presented out of band, almost as a given. The algorithmic problem is not to be questioned, problematized or investigated outside of its calculative dimensions – it is to be solved, almost like a puzzle. But how is the problem defined? How will the solution be recognized? How do the shape and content of inputs and outputs change as conceptions of the problem and the solution are updated as the algorithm is implemented? These questions of process rather than simply processing push at the procedural definition of the algorithm, and invite us to consider how algorithms are positioned in larger sociotechnical settings that include material constraints such as energy, space and time, as well as the goals and politics of people, groups and organizations.

Seaver argues that algorithmic systems aren't simply black boxes, or sites that can be opened, analyzed and understood (Seaver, 2017). The study of algorithms requires a methodological approach that recognizes how algorithms are deployed in particular social settings, as part of specific material practices, that operate in the world as part of *culture*, in addition to (not in place of) their manifestation as computational processes. In a memorable turn of phrase Seaver advises, "If you cannot see a human in the loop, you just need to look for a bigger loop."

Algorithmic systems can be quite difficult to study because they don't live in a particular place, and often aren't known by a single individual. Algorithmic processes

can be distributed between teams, systems and workflows that combine computation with people and their lived experiences. For Seaver the study of algorithmic systems in all these dimensions (including the computational) requires an approach that recognizes algorithms *as* culture, rather than algorithms operating simply *in* culture, both of which point to the use of *ethnography* as both method and theoretical orientation.

3.2. Studying Up

Ethnography usually involves some form of participant observation of people in a particular setting in order to understand social worlds, using data collection tools such as field notes, interviews and surveys. However the study of algorithmic systems is often challenged by practical barriers to data collection using traditional participant observation techniques. Attempts to understand algorithmic processes often put the researcher right into the beating heart of an organization, where information can often be guarded for competitive reasons, or because the information itself could allow the company's services to be subverted, gamed or otherwise critiqued. Technical jargon and expert knowledge distributed across individuals in an organization can act as a foil for understanding the dimensions of algorithms. The ethnographer finds themself in a position of what Nader calls "studying up", where the researcher is disadvantaged as they attempt to access a site of power (Nader, 1972). For these reasons Seaver suggests that researchers employ Gusterson's method of *polymorphous engagement*, or "interacting with informants across a number of dispersed sites … collecting data eclectically from a disparate array of sources in many different ways" (Gusterson, 1997). Seaver refers to this using the shorthand of "scavenging" which is also echoed by Kitchin who suggests ethnography for the study of algorithms as sociotechnical *assemblages* using (by necessity) a wide variety of sources:

> Interviews and ethnographies of coding projects, and the wider institutional apparatus surrounding them (e.g., management and institutional collaboration), start to produce such knowledge, but they need to be supplemented with other approaches, such as a discursive analysis of company documents, promotional/industry material, procurement tenders and legal and standards frameworks; attending trade fairs and other inter-company interactions; examining the practices, structures and behaviour of institutions; and documenting the biographies of key actors and the histories of projects. Kitchin (2016)

Ethnography of algorithmic systems is also challenged by the nature of observational data that the researcher encounters. Algorithms are created by people, but they are also performed as computational systems that, as our previous definition made clear, are a function of a particular set of inputs and outputs. These inputs and outputs happen in time and leave traces of their operation: be they transaction logs, database entries, status messages and the like. Geiger and Ribes method of *trace ethnography* is useful in these settings because it assists in "revealing the often invisible infrastructure that underlie routinized activities, allowing researchers to generate highly empirical accounts of network-level phenomena without having to be present at every node" (Geiger & Ribes, 2011). This opens the ethnographer up to studying data traces in files and databases, much as they might also study participants use of language.

3.3. Relations All the Way Down

Before diving into the specific details of how I conducted this study on appraisal in web archives it is worth briefly situating Critical Algorithm Studies in the broader tradition in Science and Technology Studies, which will be helpful for grounding my own research project.

As I have just outlined, algorithms operate as discrete computational processes that transform input into output; but at the same time they also function as part of an information *infrastructure*, in which the algorithms are defined in terms of the practices and relations they are embedded in. Susan Leigh Star, one key theorist of Infrastructure Studies, stresses that infrastructures are *relational*, that they are not constituted by a particular set of objects or artifacts so much as they are embedded in practices that happen in time:

> … we hold that infrastructure is a fundamentally relational concept. It becomes infrastructure in relation to organized practices. Within a given cultural context, the cook considers the water system a piece of working infrastructure integral to making dinner; for the city planner, it becomes a variable in a complex equation. Thus we ask, when–not what–is an infrastructure. (Star & Ruhleder, 1996, p. 4)

This attention to relations in infrastructure that Star identifies are a concern of Actor Network Theory (ANT) (Latour, 2005), or material semiotics more generally (Law (2009)). For Latour, ANT "define[s] the social not as a special domain, a specific realm, or a particular sort of thing, but only as a very peculiar movement of re-association and reassembling" (p. 7). These movements are identified by a broadening of the types of actors that can participate in relations to include so called non-humans as *actants*, which multiplies the types of and complexity of relations. Latour uses the term *symmetry* as a shorthand for this idea of granting both humans and non-humans with measures of agency. These relations are made durable in material, strategy, discourse, and performance, which allows them to persist in time, or as Latour puts it "technology is society made durable" (Latour, 1990).

In addition to tracing what is, and how these networks of relations are made durable, it is also critically important to consider the point of departure that is chosen for these

descriptions: to factor in the role of power in whose stories we tell, and to account for how things "could have been otherwise" (Star, 1990). Foucault's analysis of the micro-physics of power and governmentality clearly connects here. Yet these are questions not only of politics, but of ontology, recognizing that ontology can function as a totalizing force, but also exists as what Mol calls a *multiplicity*, where distinct practices generate new ways of being that coordinate in time (Mol, 2002).

3.4. Research Design

With these theoretical and methodological foundations in mind it is helpful to now return to my motivating research questions before diving into how I plan to answer them. The general question of how appraisal in web archives is happening actually cleaves into two interrelated sub-questions:

RQ1: How is appraisal being enacted in web archives?

RQ2: How do definitions of what constitutes *web archives* relate to the practice of appraisal?

RQ1 is *empirical* since it involves observation and data gathering to describe how appraisal is being performed in the field. The emphasis on *performance* here is intentional, since my purpose is not to make generalized and valid claims about of *all* appraisal practices in web archives. Instead I want to gain qualitative insight into the factors that motivate a discrete set of archivists in their decisions to collect content from the web, in order to better understand how appraisal is functioning in web archives.

RQ2 is *ontological* because it recognizes that decisions of what to collect from the web fundamentally define what a web archive is, and what it will become. At the same time, RQ2 also allows for appraisal decisions to be shaped by the material dimensions of web archives, or what web archives *are*. To borrow a term

from Jasanoff, the infrastructures of web archives and their affordances could (we will see) be *coproduced* by practices of appraisal (Jasanoff, 2006). For example, standards can constrain and promulgate particular types appraisal practices in web archives. But these standards in turn reflect decisions made in the design of web archiving systems. Rather than one being a cause of the other, design decisions about how to assemble software, hardware, algorithms and computational resources in web archives are themselves a form of appraisal practice.

Law & Lien (2012) refer to this dual lens expressed by my two research questions as *empirical ontology*, where questions of ontology are not concerned with describing a cosmology, uni-verse, or single dominant reality; but rather aim to describe how practices, or networks of actors, generate world views or multiple *ontologies*. As previously discussed above, the rich descriptions generated by ethnography are particularly well suited to the task of exploring these questions of practice, especially in the context of algorithms and data as culture.

To answer my research questions I conducted three empirical studies, which I will detail in chapters 4, 5 and 6. These three studies are coordinated using between-method triangulation in order to generate a thick description of appraisal in web archives [Denzin (1978); p. 302]. The three qualitative methods that will be triangulated are: thematic analysis, critical discourse analysis and ethnography. The application of each of these methods will be discussed more below, and in detail in their respective chapters. It is important to stress here at the outset that the purpose of triangulation is not for validation or verification, where the results of one method verify the results of another method, but rather to render a thick description. Taken together the data generated by these three methods is summative: one method's strength compensates for anothers blind spot.

Thick description is a term first used by linguistic philosopher Gilbert Ryle to de-

note descriptions of social phenomena that include contextual information which document participants own understanding of their motivations (Ryle, 1968). The anthropologist Clifford Geertz popularized thick description as a way to characterize successful ethnographic work, which recognizes that "what we call data are our own constructions of other people's constructions of what they and their compatriots are up to" (Geertz, 1973, p. 9). For Geertz, thick descriptive accounts of culture interpret the flow of social practices, and document enough of the context so that the account can be understood by someone who is removed (in space or time) from the site of investigation. Thick description clearly connects with the previously discussed approaches of Seaver and Gusterson which seek to investigate computational systems as culture.

The combination of the three methods I have chosen provide a critical zooming function, where the practices of multiple informants at multiple sites are analyzed, before diving into a year long field study at a specific location, in order to reassemble the relations that are discovered by zooming out again (Nicolini, 2009). The outline of these studies is as follows.

In Chapter 4, *Bots, Seeds and People* I discuss the results of Study 1, where I conducted a series of semi-structured interviews with practicing archivists, and archives adjacent actors, in order to hear how they decide how to archive regions of the web. Interviews were recorded, transcribed, and coded using *inductive thematic analysis* (Braun & Clarke, 2006) to derive key factors that influence and shape appraisal practices in web archives.

Chapter 5, Appraisal Talk, describes the results of Study 2, which applies critical discourse analysis (CDA) to a subset of the transcripts created in Study 1. CDA offers a theoretical framework grounded in critical theory for analyzing how participants' use of language reflects identity formation, figured worlds and social rela-

tions, while also addressing the larger sociotechnical context in which practice takes place.

Finally, in Chapter 6. Seeing Software I relate findings from a year long field study at the National Software Reference Library (NSRL) at the National Institute for Standards and Technology (NIST) in Gaithersburg, Maryland. Since 1999 the NSRL has maintained one of the largest collections of software in the world. While not a web archive in name, the NSRL contains software packages that have been actively collected from the web as data. Crucially, the NSRL needs to decide what software to collect from the web, and in doing so must balance the needs of their customers with the architectural constraints of their repository infrastructure, staff and funding mandate.

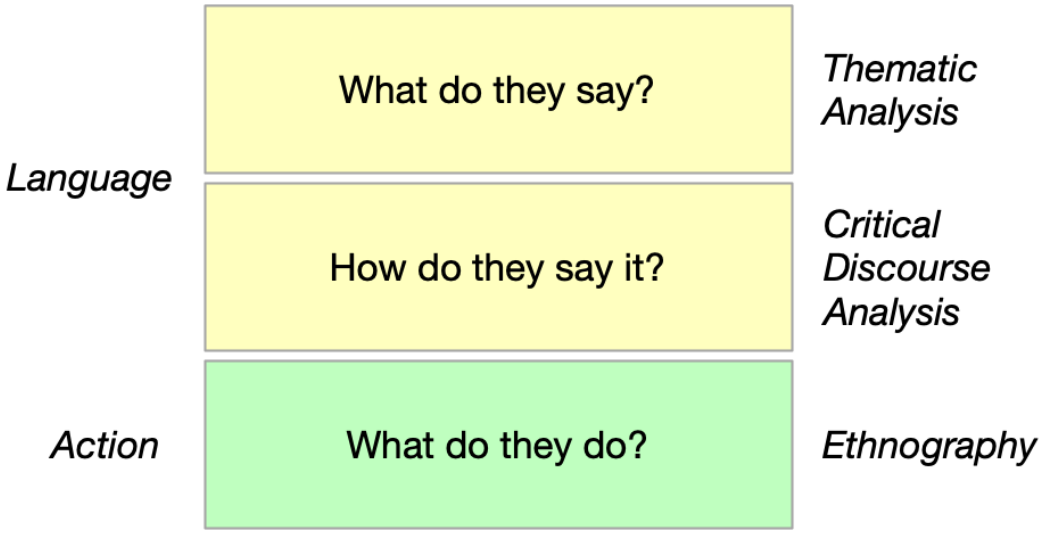

Figure 2: Between-Method Triangulation

These three studies provide a zooming function in that they start with 1) what people talk about when they talk about appraisal the web, 2) how they talk about what appraising the web, and 3) what they do when appraising the web. Conceptually 1 and 2 are about inferring practice through language use, and 3) is concerned with observed practice in the field.

It was tempting to think that these studies as a whole would provide insights about the design of web archiving systems. Especially for the NSRL study, I wanted to yield results that could potentially benefit the organizationthat was kind enough to host me. However it is not explicit purpose of my research project to derive new designs for web archiving systems, but to generate new knowledge about them. My goal in studying how appraisal is being enacted in web archives is to shed light on how web archives themselves are conceived, and how a more enlarged and theorized concept of web archiving practice can enrich information studies pedagogy and historiography. As Dourish notes in his influential piece critiquing the role of ethnography in design:

> It is practice that gives form and meaning to technology; the focus of ethnography is the ways in which practice brings technology into being. From this perspective … we might suggest that what ethnography problematizes is not the setting of everyday practice, but the practice of design … What matters is not simply what those implications are; what matters is why, and how they were arrived at, and what kinds of intellectual (and moral and political) commitments they embody, and what kinds of models they reflect. (Dourish, 2006)

4. Bots, Seeds and People

Given its vastness, volume of content, and the nature of online media, capturing and archiving the web necessarily relies on digital tools. These archiving tools typically require archivists to supply lists of website URLs or *seed lists* that are deemed important to capture. These lists are essentially a series of starting points for a web crawler to begin collecting material. The lists are managed by web archiving software platforms which then deploy web crawlers or bots that start at a seed URL and begin to wander outwards into the web by following hyperlinks.

Along with the seeds archivists also supply *scopes* to these systems that define how far to crawl outwards from that seed URL–since the limits of a given website can extend outwards into the larger space of the web, and it is understandably desirable for the bot not to try to archive the entire web. Different web archiving systems embody different sets of algorithms and as platforms they offer varying degrees of insight into their internal operation.

In some ways this increasing reliance on algorithmic systems represents a relinquishing of archival agency to automated agents and processes that are responsible for the mundane activity of fetching and storing content. The collaborative moment in which the archivist and the archival software platform and automated agents work together has not been closely examined. In this chapter I focus on answering the question of how archivists interact with web archiving systems, and collaborate with automated agents when deciding what to collect from the web.

The study of algorithms and their social effects is a rapidly growing area of research which offers multiple modes of analysis for the study of web archives (Gillespie & Seaver, 2015). Kitchin provides a review of this literature while presenting a framework of methodological approaches for the study of algorithms Kitchin (2016). He stresses why thinking critically about algorithms is so important, which is especially

relevant for thinking about the socio-political dimension of web archives:

> Just as algorithms are not neutral, impartial expressions of knowledge, their work is not impassive and apolitical. Algorithms search, collate, sort, categorise, group, match, analyse, profile, model, simulate, visualize and regulate people, processes and places. They shape how we understand the world and they do work in and make the world through their execution as software, with profound consequences.

This first study is specifically focused on how archivists interact with web archiving systems as they select material, to gain insight into how content is selected for preservation. We know that lists of URLs, or seed lists, are created, since web archiving technologies require them in order to function. But how URLs end up on these lists is not well understood. These seed lists are singular artifacts of the intent to archive, which makes them valuable excavation sites for deepened understanding of the day to day process of appraisal in web archives.

4.1. Methodology

To gain insight into how archivists are appraising content we conducted a series of semi-structured interviews with a carefully selected group of individuals involved in the selection of web content to explore and excavate these seed lists as sites of appraisal practice. Rather than providing a statistically representative or generalizable picture, the goal was to evoke a thick description of how practitioners enact appraisal in their particular work environments.

Archival appraisal is a socially constructed activity that necessarily involves the individual archivist, the organization in which they work, and the broader society and culture in which they live. Consequently the interviews did not serve as windows onto the appraisal process so much as they provided insight into what archivists talk

about their work with web archives, specifically with regards to their selection of web content for archiving or long-term preservation (Holstein & Gubrium, 2011).

Interview subjects were selected using purposeful sampling that primarily followed a pattern of stratified sampling, where both typical cases and extreme cases were selected. Typical cases included self-identified archivists involved in traditional web archiving projects at libraries, archives and museums, many of whom were on the list of attendees at the Web Archives conference in Ann Arbor, Michigan on November 12-13, 2015. The study also involved participation from extreme or deviant cases that include individuals who do not necessarily identify as archivists but are nevertheless involved in web archiving, such as researchers, local government employees, volunteers, social activists, and business entrepreneurs.

To avoid oversampling from the users of Archive-It (currently the leading service provider of web archiving services in the United States), we also recruited customers of other web archiving service providers, such as Hanzo, ArchiveSocial, and members of the ArchiveTeam community. The organization types identified in the NDSA survey results (Bailey, 2013) provided a good basis for sampling and recruitment. However, our own personal familiarity with the small but growing field of web archiving also informed the development of our participant list.

In some instances I relied on snowball sampling to recruit interview participants. There were occasions when the interview subject was not involved directly in the selection of content for their web archives. In those cases, we asked if they could refer someone that was more involved in the actual selection process. Other names were often mentioned during the interview, and if I felt those individuals could add a useful dimension to the interview I asked for their contact information.

The study recruited 39 individuals (21 female, 18 male), 27 (13 female, 14 male) of which agreed to be interviewed. A table summarizing the organization types, oc-

cupations, and roles for the interview subjects is included below. It also includes a designation of whether they were considered extreme or deviant cases (participants who do not identify themselves as archivists but are involved in web archiving duties). The tables illustrate how the our study explicitly focuses on archivists involved in the selection of web content in a university setting. Deviant cases such as the role of researcher or developer provide an opening for future work in this area.

Organization	Deviant	n=27
University	N	19
Non-Profit	Y	3
Museum	N	2
Government	N	1
Public Library	N	1
Activist	Y	1

Occupation	Deviant	n=27
Archivist	N	17
Manager	N	7
Developer	Y	2
Professor	Y	1

Role	Deviant	n=27
Selector	N	17
Technician	N	6
Developer	Y	3
Researcher	Y	1

Each interview lasted approximately an hour and was allowed to develop organically as a conversation. The interview protocol in Appendix A guided the conversation, and provided a set of questions to return to when necessary. This protocol was particularly useful for getting each interview started: describing the purpose of the study and the reason for contacting them. The interview subjects were then

asked to describe their work in web archives, and about their own personal story of how they had come to that work. After this general introduction and discussion, the conversation developed by asking follow on questions about their work and history. The ensuing conversation normally touched on the interview questions from the protocol in the process of inquiring about their particular work practices and experiences. Towards the end of the interview, the interview protocol was also useful in identifying any areas that had not been covered.

Interviews were conducted via Skype and recorded. Each participant provided informed consent via email. Participants were located in places all across the United States, so in person interviews were impractical. Because of the nature of the study the risks to participants was low, all interviews were kept confidential and all recordings and transcripts would be destroyed after the completion of the study. Consequently, we use pseudonyms to refer to our respondents and the names of their respective organizations have been obscured.

While this study was not conducted in the field over an extended period of time, it was deeply informed by ethnographic practices of memoing and fieldnote taking. These techniques were selected to document the conversation itself but also to reflect on our involvement and participation in the web archiving community (Emerson et al., 2011). During the remote interviews memos or jottings were essential for noting particular moments or insights during the interview. In some cases, these jottings were useful in highlighting points of interest during the interview itself. Immediately after each interview, these jottings prompted more reflective fieldnotes that described notable things that came up in the interviews. Particular attention was paid to themes that reoccurred from previous interviews, and new phenomena. As the interviews proceeded, a file of general reflections helped determine recurring themes and scenarios as well as unusual cases that encountered.

The process of inductive thematic analysis performed in this study relied on the use of field notes and personal memos (Braun & Clarke, 2006). The analysis began by reading all the field notes together, and then returning to do line by line coding. While coding was done without reference to an explicit theoretical framework, it was guided by our own interest in appraisal theory as a sociotechnical system that entangles the archivist with the material of the web and automated agents. Interviewee responses that specifically mentioned the selection of particular web content, and the tools and collaborations they used to enact that selection were followed up on and explored through open discussion. This analysis yielded a set of themes that will now be described.

4.2. Findings

This study reveals that web archiving involves a variety of technical and resource constraints that go beyond what is normally considered in archival appraisal theory. Archival scholars typically characterize archival selection as a process whereby human actors (archivists) primarily follow prescribed sets of rules (institutional policies and professional expectations) to accomplish the task of appraisal and selection (Boles & Young, 1985; Cook, 2011; Couture, 2005; Cox, 2010; Eastwood, 1992; Greene, 1998; Trace, 2010). This traditional notion does not adequately describe how selection occurs in the web archiving context. Instead, we found that automated agents often serve as collaborators that act in concert with the archivist. Indeed, these agents themselves are often the embodiment of rules or heuristics for appraisal. In this section, I report how crawl modalities, information structures, and tools play a significant role in selection decisions. I also highlight how resource constraints as well as moments of breakdown work to shape appraisal practice.

4.2.1. Crawl Modalities

While often guided by archivists in some fashion, the work of archiving is mostly achieved through a partnership with automated agents (bots) that do the monotonous, mostly hidden work of collecting web pages. This work includes fetching the individual HTML for web pages, and then fetching any referenced CSS, JavaScript, image or video files that are required for the page to render. Once a given page has been archived the bot then analyzes the resource for links out into the web, and decides whether to follow those links to archive them as well. This process continues recursively until the bot is unable to identify new content that the archivist has selected, is told to stop, or terminates because of an unforeseen error. Participants reflected on this process by talking about the paths that they took, often *with* their automated agents, through the web in different ways, or modalities: domain crawls, website crawls, topical crawls, event based crawls, document crawls.

In domain crawls, a particular DNS name was identified, such as *example.edu* and the crawler was instructed to fetch all the web content at that domain. These instructions often included scoping rules that either allowed the crawler to pull in embedded content from another domain, such as video content from YouTube, and also to exclude portions of the domain, such as very large data repositories or so called browser traps such as calendars that created an infinite space for the crawler to get lost in.

Website crawls are similar in principle to domain crawls, but rather than collecting all the content at a particular domain they are focused on content from a specific DNS host such as *www.example.edu*, or even a portion of the content made available by that webserver such as *http://www.example.edu/website/*.

In topic based crawls the archivist was interested in collecting web material in a specific topical area, such as “fracking” or a particular “classical music composer”.

In order to do this type of crawl the archivist must first identify the domains or websites for that topic area. Once a website or domain has been identified the archivist is able to then instruct the crawler to collect that material.

Event based crawls are similar to topic based crawls but rather than being oriented around a particular topic they are concerned with an event anchored at a particular time and place such the Deepwater Horizon Oil Spill in the Gulf of Mexico. Just as in topic based crawls, host names or website URLs must be identified first before an automated agent can be given instructions about what content to collect. With event based crawls crawling tends to extend over a particular period of time in which the event is unfolding.

In document crawls the archivist has a known web resource that they want to collect and add to their archive. This may require an automated agent of some kind, such as Webrecorder, but also could be a more manual process where an archivist collects a PDF of a report from a website and individually deposits it into their archive.

These crawl modalities were often used together, in multiple directions, by human and non-human agents either working together or separately. For example in a topical based crawl for fracking related material one archivist engaged in a discovery process of searching the web using Google and then following links laterally outwards onto social media sites and blogs. Once a set of URLs was acquired they were assembled into a seed list and given to the Archive-It service to collect. Archive-It is a subscription service run by the Internet Archive which allows cultural heritage organizations to build web archives using the infrastructure at the Internet Archive rather than locally in their organization.

Similarly when an archivist instructed Archive-It to perform a domain crawl for a large art museum, the resulting data set was deemed too large and incomplete. A proliferation of subdomains, and a multiplicity of content management systems

made it difficult to determine the completeness of the crawl. In this case the archivist used Achive-It's crawl reports as well as searching/browsing the website to build a list of particular sub-websites within the domain that were desirable to archive. Many of these sub-websites were in fact different computer systems underneath, with their own technical challenges for the web archiving bot. The larger problem of archiving the entire domain was made more feasible by focusing on websites discovered doing a failed domain crawl. This list of websites was then given to Archive-It in the form of a seed list.

4.2.2. Information Structures

In addition to the types of crawling being performed, the activities of archivists and automated agents were informed by information structures on and off the web. Primary among these structures encountered in the interviews were hierarchies, networks, streams and lists. Hierarchies of information were mentioned many times, but not by name, especially when an archivist was engaged with collecting the web content of a particular organization: e.g. the web content from a particular university or government agency. This process often involved the use of an organizational chart or directory listing the components and subcomponents of the entity in question. One participant talked about how they used their university's A-Z listing of departments as a way to build a list of seeds to give to Archive-It. In another example a government documents librarian used the organizational chart of San Mateo local government to locate web properties that were in need of archiving.

Not all web archiving projects are fortunate enough to have an explicit hierarchical map. Many appraisal activities involve interacting with and discovering networks of resources, that extend and cut across across organizational and individual boundaries. For example when Vanessa (all names are pseudonyms) was archiving web content related to the Occupy social movement she saw her organization's interest

in collecting this content fold into her own participation as an activist. This enfolding of interest and participation was evident in the network structure of the web where her personal social media connections connected her to potential donors of web content.

> It was part of that same sort of ecosystem of networks. It became clear to me through that process how important that network is becoming in collecting social movements moving forward. It was interesting watching people who had been doing collecting for decades in activist networks that they were a part of, and then these new activist networks…there wasn't a whole lot of overlap between them, and where there was overlap there was often tension. Unions really wanted in on Occupy and young people were a little bit wary of that. So social media networks became really important.

In another example a network of vendor supplied art agents supplied a museum with gallery catalogs, which were then used to identify gallery and artist websites of research value. Physical networks of agents, artists and galleries undergirded the networks of discovered websites. Indeed this particular museum used multiple vendors to perform this activity.

Another information structure that participants described as part of their appraisal process was "information streams." Information streams are content flows on the web that can be tapped into and used for the selection of content for a web archive. For example Roger who worked for a non-profit web archive described how they developed a piece of software to use a sample of the Twitter firehose to identify web resources that are being discussed. Roger also described how edit activity on Wikipedia involving the addition of external links was used to identify web resources in need of archiving. Nelson who worked as a software developer for

another volunteer organization described how he used RSS feeds to identify new news content that was in need of archiving. More traditional streams of content in the form of mailing lists and local radio, pushed content to several archivists. These streams were analyzed for reference to people, organizations and documents to seek out on the web.

While they are a bit more abstract, participants also described interacting with lists of information. The most common example of this was lists of URLs in the host reports from the Archive-It service which allowed archivists to review what host names were and were not present in their capture of web content. For example Dorothy who was collecting her university's domain:

> I definitely remember there was a lot of trial and error. Because there's kind of two parts. One of them is blocking all those extraneous URLs, and there were also a lot of URLs that are on the example.edu domain that are basically junk. Like when sometimes Archive-It hits a calendar it goes into an infinite loop trying to grab all the pages from the calendar. So what I would typically do is look at the list of the URLs. Once you've done a crawl, a real crawl, or a test crawl that doesn't actually capture any data, there's this report that has a list of hosts, for example facebook.com, twitter.com and then next to that there's a column called URLs and if you click the link you get a file, or if it's small enough a web page that lists all the URLs on that domain. So one thing that I would try to do is visually inspect the list and notice if there's a lot of junk URLs.

The question of what is and what is not junk is the central question facing the archivist when they attempt to archive the web. The reports that Archive-It provides at the host name level are an indicator of whether the crawl is missing or in-

cluding things that it should not. Scanning lists of host names and URLs happened iteratively as multiple crawls were performed.

When considering how participants talked about these hierarchies, networks, streams and lists of information it became clear that they were traversing these structures themselves using their browser, as well as instructing and helping the archival bots do the same. The domain knowledge of the archivist was a necessary component in this activity, as was the ability for the bot to rapidly perform and report on highly repetitive tasks.

4.2.3. Time and Money

Another thematic feature that emerged from the fieldnotes around the interviews were the material constraints of time and money in the human-machine collaboration of web archiving. Time and money are combined here because of the way they abstract, commensurate and make appraisal practices legible.

Many web archiving projects cited the importance of grant money in establishing web archiving programs. These grants often were focused on building technical capacity for web archiving, which itself is not directly tied to the appraisal process. However it is clear that the technical ability to archive web content is a key ingredient to performing it. Grant money was also used to archive particular types of web content. For example, one university used grant money to archive music related web content, and another university received a grant to focus on state government resources.

The most common way that money was talked about by participants was in subscription fees for web archiving services. Archive-It subscribers pay an annual subscription fee to archive web content. The primary metric of payment is the amount of data collected in a given year. Interviewees often mentioned that their ability to

crawl content was informed by their storage budget. In one example an archivist set the scoping rules for a full domain crawl of her university such that software version control systems were ignored because of the impact it was having on their storage allocation. Dorothy, who was a user of the ArchiveSocial service needed to reduce the number of local government social media accounts that it was archiving because her subscription only allowed a certain number of accounts to be collected.

Time manifested in the appraisal of web content at human and machine scales. In one common pattern, archivists set aside time every week, be it a day, or a few hours, for work on the discovery of web content. In one case, Wendy set aside time to read filtered emails about local news stories. In Lisa's case, she set aside a meeting time every week for her acquisition team to get together and review potential web sites for archiving by inspecting websites together on a large screen monitor.

Time was also evident in the functioning of automated agents, because their activity was often constrained and parameterized by time. For example archivists talked about running test crawls in Archive-It for no longer than 3 days. Dorothy talked about the information being gathered in near real time from social media accounts that Archive-It was monitoring:

> The archiving is by the minute. So if I post something, and then edit it in five minutes then it is archived again. If someone comments on something and then another person comments it is archived again. You don't miss anything. A lot of the other archiving companies that we've talked to say they archive a certain number of times a day: maybe they archive at noon, and at 5, and at midnight, and there's an opportunity to miss things that people deleted or hid.

In this case the software was always on, or at least appeared to be always on at human time scales. The web content itself also had a time dimension that affected

appraisal decisions. For example the perceived cumulativeness of a website was an indicator of whether or how often material was in need of archiving. Blogs, in particular, were given as examples of websites that might need to be crawled less because of the ways that they accumulated, and did not remove content.

Another motivation for linking time and money in this way is because of how they entail each other. The time spent by archivists in discovery and evaluation of web content for archiving often has a monetary value in terms of salary or hourly wages. Similarly the amount of time spent crawling is often a function of the amount of data acquired, and the cost for storage.

4.2.4. People

One might assume that the work to appraise web archives necessarily involves archivists. However, the interview data made it clear that not all the people involved in appraisal called themselves archivists, and they often worked together with human and non-human agents in collaborative relationships that extended beyond the archives itself.

At one large university archives, a series of individuals were involved in the establishment of their web archives. Their effort extended over a 15-year period that started with Kate who pioneered the initial work that ultimately led to a mainstreaming of web content into the archives. Multiple staff members, including Jack and Deb who were field archivists responsible for outreach into the university community, and around the state. The field archivists selected web content, which was communicated to Phillip, another archivist, who managed their Archive-It subscription, performed crawls and quality assurance. Jack and Deb actively sought out records in their communities by interviewing potential donors, to determine what types of physical and electronic records were valuable.

John worked as a software developer for a volunteer organization that performed focused collecting of web content that was in danger of being removed from the web. He was a physics student who was interested in using his software development skills to help save at risk web content. John collaborated with 20-30 other volunteers, one of whom is Jane who worked at a large public web archive, and was routinely contacted via email and social media when websites were in danger of disappearing.

Many interviewees reflected on their own participation in the activities and events that they were documenting. Recall Vanessa who was working to collect web content related to the Occupy movement. She and her colleagues at the library worked to document the meetings and protests from within the movement itself. One of her colleagues worked on the minutes working group which recorded and made available the proceedings of the meetings. In another case two archivists and separate institutions were working together to document the use of fracking in their respective geographic areas. They worked together to partition the space as best they could by region, but many businesses and activist organizations worked across the regions.

While collaboration across organizational boundaries was evident, several participants noted that duplication of web content was not widely viewed as something to be avoided. Many commented that duplication was one way to ensure preservation, following "lots of copies keeps stuff safe" (LOCKSS). Local copies of resources that are available elsewhere can be of benefit when using the data:

> If I can't get a copy it doesn't exist in the same way. I think that there is still a lot to being able to locally curate and manage collections and the fact that it's over in another space limits, or puts some limits on the things that can be done with the data now and in the future. Sure right now I've got a great relationship with a guy that knows how to get the stuff. But what happens in five years when those relationships end?

How do our students and researchers get access to the data then?

In addition the locus of web archiving work shifted within organizations from one department to the other as key individuals left the library, and as web content was migrated from one system to another. This turbulence was common, especially in the use of fellowships and other temporary positions.

4.2.5. Tools

We have already discussed some tools of the trade that archivists use for collecting websites: the Internet Archive, Archive-It, ArchiveSocial and Hanzo are notable ones that came up during the interviews. These tools are really more like services, or assemblages of individual tools and people interacting in complex and multilayered ways. An investigation of each of these services could be a research study in themselves. These tools largely require intervention by a person who guides the tool to archive a particular website, or set of web resources using a seed list or the equivalent. Rather than dig into the particular systems themselves it is useful to attend to the ways which tools were used to fill in the gaps between these platforms and their users.

Consider the ways in which spreadsheets were used almost ubiquitously by interviewees. These spreadsheets were occasionally used by individuals in relative isolation, but were most often used to collaboratively collect potential websites that were of interest. Google Sheets in particular allowed individuals to share lists of URLs and information about the websites. Archivists would share read-only or edit level permissions for their spreadsheets to let each other know what was being collected. These spreadsheets were later transferred into a web archiving service like Archive-It as seed lists. In the process much of the additional information, or provenance metadata concerned with the selection of a website was lost in translation.

Often times web forms of various kinds were used as front ends on these spreadsheets. These forms mediated access to the spreadsheets and provided additional guidance on what sorts of data were required for nominating a web resource for the archive. Tracy developed a custom application for tracking nominations, so different parties could see each other's nominations. Tracy noted that one of its drawbacks was that the tool did not link to the archived web content when it was acquired.

Email was also widely used as a communication tool between selectors of websites and the individuals performing the web crawling. In one case a technician would receive requests to crawl a particular website via email, which would initiate a conversation between the technician and the selector to determine what parts of the website to archive. This process would often involve the technician in running test crawls to see what problem areas there were. Several archivists spoke about how they subscribed to specific local news aggregators that collected news stories of interest.

However, email was not the only communication method used in the appraisal process. As already noted social media, particularly Twitter, was used as a way of communicating with prominent web archiving individuals when websites were in need of archiving. In one case IRC chat was also a way for volunteers to talk about websites that were in need of archiving, and to coordinate work. These conversations were extremely important because they embody the process of determining the value of resources.

Many interviewees used the Archive-It service and commented on the utility of test crawls. Test crawls were essentially experiments where the archivist instructed the crawler to archive a particular URL using particular scoping URLs to control how far the crawl proceeded. Once the crawl was completed the archivist would examine the results by browsing the content and comparing to the live website. The archivist

would also examine reports to look at the amount of data used, URLs that were discovered but not crawled either because of time or because they were blocked by the scope rules. The experiments were iterative in that the results of one test would often lead to another refined test until the crawl was deemed good. Almost all participants talked about this process as quality assurance or QA instead of appraisal, despite the fact that it was ultimately a question of what would and would not go into the archive. One exception to that rule was an archivist who had 10 years experience doing web archiving with multiple systems who referred to this as pre-crawl and post-crawl appraisal.

It is notable to observe how engineering terminology like quality assurance has crept into the language of the archive where appraisal would be a more apt term. One archivist also noted how archival notions of processing and appraisal which are normally thought of as distinct archival activities get folded together or entangled in the process of test crawling. Indeed one participant went so far as to say that the process of web archiving actually felt more like collection building than archiving.

In few cases, the Domain Name Service itself was used as a service to discover subdomains that were part of a university's domain. A large number of target hostnames were discovered, which were then prioritized in order to build a seed list. In another case knowledge of the rules around the .mil DNS top level domain were used to determine websites of interest for archiving government sites. However these rules were imperfect as some US government websites would use the .com top level domain, such as US Postal Service.

Another prominent technology that participants mentioned was content management systems. In many cases archivists had experience working as web designers or information architects. They had used content management systems like Drupal, Ruby on Rails, WordPress, etc. The archivist would use this knowledge to decide how to

crawl websites and diagnose problems when they arose.

4.2.6. Breakdown

One of the more salient findings during analysis was the locus of breakdown which made the relations between people, tools, and web infrastructure more legible. These moments of breakdown also lead to greater understanding of how the tools operated, and generated opportunities for repair and innovation (Steven J. Jackson, 2014).

Charles was attempting to do a full domain crawl of his university's domain with the Archive-It tool. An unfortunate side effect of running this crawl was that portions of the university website were put under more significant load than usual, became unresponsive, and crashed. IT specialists that were responsible for these systems incorrectly identified the crawlers as a denial of service attack, and traced them to Archive-It. An email conversation between the technicians and Archive-It led to the technicians at the university connecting up with the archivists who were attempting to archive web content–at the same institution. This situation led to lines of communication being opened between the library and the central IT which were not previously available. It also led to increased understanding of the server infrastructure at the university which was housed in four different locations. The IT department became aware of the efforts to archive the university's web spaces, and began to notify the archivist when particular websites were going to be redesigned or shutdown and in need of archiving.

In another case John used a command line web crawling and archiving tool called wget to collect web content. wget was used to generated a snapshot of web content and serialize it using the WARC file format. He then used another piece of software playback tool called WebarchivePlayer to examine the data stored in the WARC file

to see how complete the archive was. In some cases he would notice missing files or content that failed to load because the browser was attempting to go out to the live web and he had disabled Internet access. This breakdown in the visual presentation of web resources would prompt John to use the browser's developer tools to look for failed HTTP requests, and trace these back to JavaScript code that was dynamically attempting to collect content from the live web. He would then also use this knowledge to craft additional rules for wget using the Lua programming language, to fetch the missing resources. When his examination of the WARC file yielded a satisfactory result the resulting Lua code and wget instructions were bundled up and deployed to a network of crawlers that collaborated to collect the website.

As previously discussed, storage costs are another point of breakdown when archivists are deciding what web content to archive. Several participants mentioned their use of test crawls in an attempt to gauge the size of a website. The full contours of a website are difficult to estimate, which makes estimating storage costs difficult as well. Some participants were able to communicate with individuals who ran the website being archived in order to determine what content to collect. Roger, who was mentioned earlier, was able to got into conversation with an engineer who worked at a video streaming service which was in the process of being sold. Together they determined that the full set of data was 1.1 petabytes in size, which (after consultation with the directory of that archive) made it very difficult to think about archiving in full.

> I went back to the developer and asked: could you give me a tally of how many videos have had 10 views, how many videos have had 100 views and how many videos have had a 1000 views? It turned out that the amount of videos that had 10 views or more was like 50-75 TB. And he told me that 50% of the videos, that is to say 500 TB had never been

> viewed. They had been absorbed and then never watched. A small amount had been watched when they were broadcast and never seen again. We had to walk away from the vast majority. Given that we can't take them all, what are the most culturally relevant ones? We grabbed mostly everything that was 10 or more. The debate is understandable. In an ideal world you'd take it all. The criteria we've tended to use is, I always like to grab the most popular things, and the first things. So if you have a video uploading site I want the first year of videos. I want to know what people did for the first year when they were faced with this because there's no questions this is very historically relevant. But I also want people to have what were the big names, what were the big things that happened. And that's not perfect.

In this case a breakdown that resulted from the size of the collection and the available storage became a site for innovation, and an opportunity to make legible appraisal decisions around what constitutes culturally significant material.

Another extremely common case of breakdown is when a robots.txt file prevented a crawler from creating a high fidelity capture of a website. A robots.txt file instructs automated agents in what resources it can and cannot request. Frequently content management systems will block access to CSS or image files which makes a web archive of the pages visibly incomplete, and difficult to use. Many (but not all) archives attempted to be polite by instructing their web archiving bots to respect these robots.txt files. When they encountered a problem they would often need to reach out to someone at the organization hosting the website. When contact was made the robots.txt file would sometimes be adjusted to allow the bot in. The archivist became aware of how the website was operating and the website owner became aware of the archiving service. In one instance this communication channel

led a website owner to make more cumulative information available on their website instead of replacing (and thus removing) older content. In some sense the website itself adapted or evolved an archival function based on the interactions between the archivist and the manager of the website being archived.

4.3. Discussion

On the one hand these research findings demonstrate a somewhat mundane but perhaps comforting finding that in many ways appraisal processes in web archives appear to be congruent with traditional notions of appraisal. We saw documentation strategies Samuels (1986) at play in many cases where a collaboration between records creators, archives and their users informed decisions about what needed to be collected from the web. The appraisal technique of functional analysis was also used by archivists as they analyzed the structure of organizations in order to determine what needed to be collected. We also saw postcustodial theory Cook (1993) in operation when archivists interacted with website owners, and in some cases encouraged them to adopt archival practices. So rather than a particular archival institution being responsible for the preservation and access to documents, the responsibility is spread outwards into the community of web publishing.

A recurring theme in the analysis above was the archivists' attention to contemporary culture and news sources. We recall one participant who spoke of her mentor, who had set an example of taking two days every week to pore over a stack of local newspapers, and clip stories that contained references to local events, people and organizations to explore as record sources. She spoke of how she continued this tradition by listening to local radio, subscribing to podcasts, RSS feeds and email discussion lists. She then regularly noted names of organizations, people and events in these streams as potential record sources. While not all interviewees spoke explicitly of this practice being handed down, the attention to local news sources was a

common theme, particularly when it came to processing information streams. This attention to current events while simple, is extraordinarily powerful, and reminiscent of German archivist Hans Booms:

> The documentary heritage should be formed according to an archival conception, historically assessed, which reflects the consciousness of the particular period for which the archives is responsible and from which the source material to be appraised is taken. (Booms, 1987, p. 105)

Echoes of Booms can also be found in this description by Roger of how his archive's appraisal policies are enacted:

> The greater vision, as I interpret it, is that we allow the drive of human culture to determine what is saved. Not to the exclusion of others, but one really good source of where things are that need to be saved is to see what human beings are conversing about and what they are interacting with online.

Websites, search engines and social media platforms are material expressions of the transformation of content into computational resources, with centers of power and influence that are new, but in many ways all too familiar. The continued challenge for archivists is to tap into these sources of information, to deconstruct, and reconstruct them in order to document society, as Booms urged. In the shift to computational resources there is an opportunity to design systems that make these collaborations between archivists, automated agents and the web legible and more understandable for all parties, and particular for the future researcher who is trying to understand why something has been archived.

The appraisal processes that are being enacted by archivists are not always adequately represented in the archive itself. Recall the spreadsheets, emails and chat

systems that are used during appraisal, that all but disappear from the documentary record. These systems are being used to fill the broken spaces or gaps in the infrastructure of web archives. Each of these hacks, or attempts at creatively patching archival technology, is a potential design hint for archival tools and platforms.

For example, if spreadsheets can be collaboratively used by a group of archivists to record why a web resource was selected, who selected it, and other administrative notes, perhaps this collaborative functionality could be incorporated into the web archiving platforms themselves? One opportunity of future work would be to examine these sites of breakdown in greater detail, in order to help archivists and their automated agents create a more usable and legible archival record. Further examination of how consensus is established when archivists are collaborating would also be a fruitful area to explore in order to understand *how* archivists are collaborating with each other using these technical systems.

Another significant theme is found in the collaborative sociotechnical environment made up of archivists, researchers (the users of the archive), and the systems/tools they use in their work. The inner workings of the archive always reflect or remediate (Bolter, 2016) the content they attempt to preserve and provide access to. As electronic records and the World Wide Web have flourished, the architecture of the archive itself has necessarily been transformed into a computerized, distributed system, whose data flows and algorithms reshape the archival imagination itself (Taylor, 1992). Even with its narrow focus on the appraisal decisions made by archivists this study demonstrates that archivists have rich and highly purposeful interactions with algorithmic systems as they do their work of selecting and processing web resources. Time and again archivists used these systems, and cleverly arrived at techniques for imagining the dimensions of the resources they were collecting, the fidelity of the representations created, and ultimately the algorithmic processes that they were di-

recting.

It's not very difficult to imagine a near future where the archival record is complete. Nothing is lost. Everything is remembered. To a few this is a big data panacea, but to most of us it is a dystopian nightmare of the panopticon. Fortunately we find ourselves somewhere in between these two unlikely extremes. This study asked a simple question of how URLs end up being selected for an archive. The findings illustrate that archivists talk about their appraisal of the web as part of a dense of network of actors that includes bots, record creators and the web instrastructures that operate in a flattened space. In the next chapter I will do a close reading of several of these interviews in order to study not only *what* they talk about when they talk about archiving the web but *how* they they talk about it.

5. Appraisal Talk in Web Archives

Archival appraisal is generally understood to be the work that archivists do to identify materials that have sufficient enduring value to justify their being cared for in an archives (Pearce-Moses & Baty, 2005). It is typical for archives to have different notions of value, and these are often expressed in the collection development policies that archivists use in the selection work they do while processing collections. Appraisal decisions continue to be made as collections are cared for and as the demands of new records impinge on the archive's ability to store them (Rapport, 1981).

While the values ascribed to individual archives differ, the activity of appraisal is central to the work of all archivists. The cumulative effect of these appraisal decisions shapes the historical record and, by extension, our knowledge about the past and our social memory (Jacobsen et al., 2013). This value-driven process of appraisal has many facets, which sometimes can seem to suffuse all of the archivist's work. For example, the values that drive appraisal also find expression in the ways archives are arranged and described, which in turn determine how they are accessed (Yakel, 2003). To describe this moment at the inception of an archive with more specificity, Eric Ketelaar coined the term *archivalization*:

> It is archivalization , a neologism which I invented, meaning the conscious or unconscious choice (determined by social and cultural factors) to consider something worth archiving. Archivalization precedes archiving. The searchlight of archivalization has to sweep the world for something to light up in the archival sense, before we proceed to register, to record, to inscribe it, in short before we archive it. (Ketelaar, 2001)

The searchlight of archivalization that Ketelaar imagines for us is the process of

deciding what to remember, no matter what material form the archive takes. In this chapter, I use *archivalization* and *appraisal* somewhat interchangeably, but I use *archivalization* to refer specifically to the initial moments in which a decision is made about what to preserve and what not to preserve. Ketelaar goes on to remind us that "technology changes the archivable". The technologies of record production that we create inevitably shape both *what* and *how* records get archived (Schwartz, 1995).

In this chapter I will explore how these expressions of archivalization, the specific moments of appraisal, are being performed in web archives in order to gain insight into how the infrastructure of the Web is shaping our attempts to preserve it. In order to investigate these moments of archivalization, I will use critical discourse analysis to closely examine how web archivists talk about their appraisal decisions. As my findings illustrate the global address space of the Web and the immediacy of its underlying protocols have occasioned a shift in the nature of appraisal, particularly with regard to the trust relationship between the documenter and the documented.

5.1. Methodology

One way of investigating the phenomenon of archivalization is to qualitatively analyze how archivists talk about their appraisal work: to look at the words they use, the conventions they have established, the context they share, the ways they learn from each other in communities of practice, and the political work that these communicative practices perform (Wenger, 1998). To address these research objectives, I undertook a critical discourse analysis (CDA) of a set of ethnographic interviews of individuals involved in selecting web content for archives. I employed CDA because it offers a theoretical framework, grounded in critical theory, for analyzing the way in which participants' use of language reflects identity formation, figured worlds, and social relations, while also addressing the larger socio-cultural context

in which practice takes place. CDA helps to examine how language use connects with issues of ideology and power, which are particularly relevant when considering archival appraisal as an inherently political act, in line with a critical archival studies' perspective (Caswell et al., 2017).

Sociolinguist James Paul Gee, a practitioner and theorist of CDA, noted that "there are solid linguistic, even grammatical grounds, on which to argue that all language-in-interaction is inherently political" (Gee, 2011). Indeed, CDA is a theoretical approach to language use rather than a method as such, and those who practice CDA bring a variety of discourse analysis methods to bear in their analyses (Wodak & Meyer, 2001). Gee's research centers on the fields of education, literacy, and media studies. This focus makes his work particularly relevant for analyzing the ways archivists talk about web archives. He elucidates *seven building tasks* that language performs to reflect and produce social relations. These building tasks involve (J. Gee, 2014, pp. 95–98):

1. **significance**: how language is used to foreground and background certain things
2. **activities**: how language is used to enact particular activities
3. **identity**: how language is used to position specific identities and make them recognizable
4. **relationships**: how language is used to construct relationships between people and things;
5. **politics**: how notions of value and norms are established in the use of language;
6. **connections**: how language is used to connect and disconnect ideas, activities, and objects; and
7. **sign systems and knowledge**: how language positions particular sign sys-

tems, or ways of knowing and believing.^56

Gee also provides a set of methodological tools that support the analysis of linguistic performances, or building tasks – in other words, tools that are used to dissect the ways that language produces social relations. For Gee, words do actual work in the world: "Whenever we speak or write, we always and simultaneously build one of seven things or seven areas of 'reality' " (J. Gee, 2014, p. 94). While I draw on several of Gee's building tasks and tools in my analysis, as I immersed myself in my transcription data, I became particularly focused on the building tasks related to *relationships*, *identity*, and *politics*.

While language is important, it is not the only means by which archivists build community in their work. CDA also allows the researcher to examine language use in relation to non-linguistic elements such as technology, infrastructure, and setting. Although this study focuses specifically on linguistic discourse, software utilities, infrastructures, and the geographic dispersion of Internet communication provide important dimensions for understanding the work of appraisal in web archives. Some of these factors emerge below in the discussion of the results. In the table below I have included a brief glossary of some notation that is used in the transcripts included here:

Notation	Meaning
//	Final intonation contour, like a period in writing
/	Non-final intonation contour, like a comma
[segment]	Overlapping talk
=	Latching: two utterances that follow one another without pause
WORD	An emphasized or stressed word

In addition to Gee's seven building tasks, I used several of his specific tools for

discourse analysis; these are mentioned in the context of the findings they helped illuminate.^62

5.2. Findings

The findings draw from interviews with three subjects, who I refer to using the pseudonyms Jim, Jack, and Carly. These excerpts, from the 30 hours of interviews coded for analysis, have been selected specifically because of the way these informants talked about their appraisal work and the manner in which they involved other people in their decision-making processes. Gee's seven building tasks, which allow language users to shape social realities, were chosen to provide a framework for insights into how and why appraisal in web archives is being performed – not to quantify or otherwise make generalized claims about the practice as a whole.

As I performed close readings of the transcripts, Gee's building task related to relationships appeared particularly useful for examining the participants' use of language. Gee explains this building task by associating it with his relationships building tool, which is used to analyze relationships found in language:

> For any communication, ask how words and various grammatical devices are being used to build and sustain or change relationships of various sorts among the speaker, other people, social groups, cultures and/or institutions. (J. P. Gee, 2014, p. 121)

Focusing on relationships present in the archivists' use of language in turn exposed two more of Gee's building tasks: those related to identities and politics. Regarding identities, Gee advises researchers to ask "what socially recognizable identity or identities the speaker is trying to enact or to get others to recognize" (J. P. Gee, 2014, p. 116). Regarding politics, Gee is primarily concerned with how language performs the distribution of "social goods" or enacts day-to-day politics rather than

with formal systems of government. The discussion of the findings below is organized around three themes that emerged when using the questions Gee suggests in relation to these three building tasks: the themes of hierarchies, mentorship, and structures.

The excerpts included below are not typical quotations because they aim to capture the stresses, pauses, and rhythms of speech using the aforementioned notation. The lines are sparse, numbered, and spatially organized to reflect the individual utterances that form the recognizable speech units that will be analyzed. The length of the excerpts varies depending on the amount of context that is useful for the discussion. The excerpts used here are not intended to be representative of archivists in general but are used to provide insights into particular factors that are at play in the appraisal of web content.

5.2.1. Hierarchies

Jim works as an archivist in a non-profit organization that does a large amount of web archiving as well as digitization. In the following excerpt, Jim describes a situation where a large amount of data was being archived from a video streaming provider that was going out of business. I was asking Jim to recall a time when he had needed to make a decision about whether or not to archive a particular website or document. The excerpt provides a particularly salient snapshot of the type of discussion that goes on in his organization when archivists decide how to archive a large amount of content from the Web:

Line	Speaker	Utterance
1	Jim	The petty disk /
2		that I have at my disposal /
3		without having to go to higher ups /

Line	Speaker	Utterance
4		is about 10 terabytes //
5	Ed	mm-hmm //
6	Jim	So if I find a job and the job is like an eight-terabyte job /
7		I don't need to bring it up with the Archives //
8		Um, if I discover that it's gonna be 30 40 50 petabytes /
9		I go to my superior /
10		Ariana Reese /
11		and Ariana initially will go /
12		"Why? /
13		Convince me this is a good idea." //
14	Ed	mm-hmm
15	Jim	And I'll say, "Well, it's cause of this this." She'll go, "Okay that sounds like a good idea. Go ahead." /
16		you know, like as a stopgap //
17		But then it will be like /
18		Well it's extremely controversial /
19		it's the stuff=and then she'll be like /
20		"Okay well /
21		if it could possibly blow back on the Archives /
22		or if we could potentially be facing some kind of issue with it /
23		let's go have a chat with Greg" /
24		and now it's me and Greg and Ariana saying /
25		"Are we gonna do this?" //
26	Ed	mm-hmm

Line	Speaker	Utterance
27	Jim	Now when I mention these /
28		I mention them like it's some sort of whatever /
29		you know we're talking /
30		I'd have to say that that's me and Greg and Ariana going /
31		whether this job /
32		has happened like /
33		six times in the last three years /
34	Ed	mmm
35	Jim	maybe? /
36		Ooom [and Ariana]
37	Ed	[and did] /
38		was Real TV /
39		was this one of the examples? /
40		the Real TV one or?
41	Jim	Well Greg helped me get in contact with the employees /
42		Greg was already on the ground with it.
43	Ed	Oh okay //
44	Jim	and Greg /
45		KNEW /
46		that it was going to be a lot of data /
47		and was like /
48		"Okay so [be a little more] /
49	Ed	[ahhhh]
50	Jim	careful with this"

As I asked Jim to recall a specific occasion when he had needed to make an appraisal decision, he recalled a situation that emphasized relationships with other employees in his organization. Note the use in line 3 of "higher ups," which invokes the concept of an organizational hierarchy or chain of command that involved his manager, Ariana, and Greg, the director of the archives. These relationships are foregrounded and frame the decision that is being made. Invoking the organizational hierarchy in this manner lends weight and formality to the appraisal decision, while also working to lead us away from a discussion of the appraisal criteria. The moment of archivalization is surfaced and then effaced. Just as the hierarchy is emphasized, the details of the actual decision-making process are elided with rapid speech and the use of "this this" in line 15 to refer to the actual appraisal criteria. We do not actually know what Jim says to Ariana to persuade her that the video content is worth saving: the "this" references are stressed through repetition, but they both lack a referent. Jim indicates that these conversations are infrequent and that the initial decision to archive this content came directly from Greg. The decision to archive this content started at the top, came down, and then went back to the top of this appraisal ladder again. The circularity and vagueness of these hierarchical relationships suggests that they could be operating as a rhetorical device to formalize what is otherwise a much less structured and more organic process. One additional relationship that is identified only near the end of this segment is the connection between Greg and Jim, inside the archives, and an employee outside at the video streaming organization. Archives often refer to these connections as donor relationships, because they broker communications between individuals or organizations that are donating materials and the receiving archives. Greg is described as helping Jim contact an employee and as already being "on the ground" with the process, which casts the archives in the role of actively seeking content and not simply receiving content. This is a highly significant donor relationship, which we will return to below.

5.2.2. Mentorship

Jack is an archivist at a large university in the United States, which he joined a few years ago after leaving a previous job as an archivist at another university. In this segment, Jack describes how he came to work on a web archive that documents the activities of the fracking industry.

Line	Speaker	Utterance
1	Jack	I came to University A actually wanting to /
2		to drive some more um /
3		I guess professional engagement around the legacy of fracking in this state //
4		I came from the University B where /
5		we had a lot of really intense collections around environmentalism /
6		and energy development in the state /
7		and it was a sort of an area of programming for the archives there //
8		And one of my close colleagues there had done a project /
9		basically sort of like anticipating the next energy boom in the state /
10		which coincidentally was fracking //
11		which also coincidentally was um something that 30 or 40 years ago /
12		a company wanted to um /
13		explore by /
14		detonating a series of underground nuclear explosions to
15		stimulate the gas um /

Line	Speaker	Utterance
16		to the surface [chuckle] //
17	Ed	Whaaat?
18	Jack	And we had some collections on on that /
19		sort of like the protest effort that um /
20		that killed that endeavour //
21		So she had done just a lot of really interesting stuff=
22		oral histories exploring the boom and bust and /
23		so I always kind of had in the back of my mind that coming /
24		to University A /
25		I had other reasons for coming /
26		but that in coming here
27		this was an issue I wanted to explore //
28		But I didn't really have an opportunity to push it /
29		until I saw / some news announcement somewhere /
30		I don't know where /
31		but University C announcing that they were /
32		going to start this project /
33		to document the fracking anti-fracking activism in the state //
34		And I immediately took it to our associate dean /
35		who at the time was Mark Dalton //

Again, we see that relationships figure prominently in this description of how decisions are made about what to archive from the Web. In lines 34–35, a hierarchical

relationship between Jack and Mark is positioned as one of the key moments of appraisal, as we saw in the previous example. We also see several organizational relationships traced between Jack's current university (A), his previous university (B), and an institutional collaborator (C). However, unlike in the previous example, these are not hierarchical relationships but links of influence and practice. The first relationship, between organization A and organization B, is mediated by Jack's own professional history. He worked at both of these organizations and mentions them in order to highlight a specific mentoring relationship between himself and a colleague at University B, who had done significant archival work around documenting fracking. His description of this relationship as "close" and as a source of inspiration lacks the clinical tone used in descriptions of the previous hierarchical relationships. This mentoring connection knits archival practices for oral histories together with those for websites, using the shared interest in documenting environmental issues and activism. There is also a relationship between Jack and an individual at University C; their shared interest in documenting fracking activated Jack's ability to begin work on the collection and also became the seed of a collaboration. It is important to note the implicit role that the Web plays in this collaboration. The distributed, globally accessible information space of the Web means that Jack and his collaborator at University C needed to partition their work geographically. Unlike physical collections, which can be in only one place at one time, the public Web is available to everyone who has a computer and an Internet connection. Jack is not stymied by University C's move to document fracking but is emboldened to participate. While a request for approval again moves up the organizational hierarchy, the initial impetus (archivalization) comes laterally, from a peer at another institution, and from the past in the form of his mentor. Rather than being a discrete event, the moment of archivalization actually involves an assemblage of actors removed in both time and space. In the following excerpt, we continue to look at mentoring relationships

as evidence of an emergent community of practice including archivists and web archivists. In this segment, we hear from Carly, who explains that she became involved in web archiving while working as an archivist for over a decade at several large research universities, where she spent a significant amount of time performing web archiving.

Line	Speaker	Utterance
1	Carly	Yeah, that's kind of how I've always /
2		Back in the day when I first started in GovDocs /
3		one of my mentors /
4		she was a local docs librarian at University D /
5		and her approach /
6		and I feel I would love to figure out a way to do this better /
7		So she actually just got the three major newspapers in the county area /
8		and she=we would pile them up for her /
9		and she would just take like a day a week /
10		and she would go through and CLIP /
11		the news articles /
12		and then she would make sure she got /
13		the documents that were mentioned um /
14		in them //

Here, Carly is connecting her practice in appraising web content with her professional experience of working as a government documents librarian. Carly specifically uses "one of my mentors" to emphasize that she learned from a specific individual – and the plural form identifies this person as one of several mentors she has

learned from during her career. References to these mentoring relationships suggest that archivalization draws on a network of learning through a community of practice. Carly's attention to specific details, such as the "three major newspapers," how her mentor "would pile them up" – as well as the stressed "clip" – recall the physical process of doing the work. The confusion in subjects in the "she=we" points to Carly's reconstruction of the scene for this appraisal work and indicates that she was one of several people working together as part of a team. It is also apparent from the repair in line 6 that Carly feels that this material process does not have a direct analogue in her current web archiving work (even though she goes on to talk about her use of email discussion lists, RSS feeds, and bookmarks later in the interview) and that it could be useful to find one.

5.2.3. Structures

There is something else going on in Carly's discourse, which may be apparent only to an outsider to the library and archives profession. To draw this out, it can be useful to follow the guidance in Gee's making strange tool: "For any communication, try to act as if you are an 'outsider.' Ask yourself: What would someone (perhaps even a Martian) find strange here (unclear, confusing, worth questioning) if that person did not share the knowledge and assumptions and make the inferences that render the communication so natural and taken-for- granted by insiders?" (J. P. Gee, 2014, p. 19) Specifically, the reference to "getting started in GovDocs" (line 2) and the expression "local docs librarian" (line 4) speak to a particular type of work that is not necessarily directly tied to the work of web archivists. Government documents librarians are trained librarians who focus on collecting, preserving, and providing access to documents published by federal, state, and local governments. This type of work came about in the United States after the establishment of the Federal Depository Library Program and the Government Printing Office by the *Printing Act*

of 1895. As such, it is highly regulated work that is guided by policy. The work of scanning the "three major newspapers" and looking for references to "documents" was being done in the context of this highly politicized activity. How were the major newspapers selected, and what factors influenced their selection? What government documents – material output of governmental activities that present a view of society from the perspective of the state – were librarians looking for? This context for the newspaper clipping and note taking is also the experience that guides Carly as she decides what to archive from the Web. The scanning of newspapers for references to government documents is a precise moment in the process of archivalization. We see this same political aspect at work more explicitly in this final excerpt from Jack, who is reflecting on his work to document fracking.

Line	Speaker	Utterance
1	Jack	I really see like one of / my next curatorial responsibilities being um /
2		not really more crawling or more selecting /
3		but using the connections I've made here /
4		to get more contact and more dialogue going with um /
5		with the actual communities I've been documenting //
6		And I'm a little nervous about how it's gonna go /
7		because I went ahead and crawled a bunch of stuff /
8		without really doing that in advance //
9		I'm also a little nervous about it because /
10		through our biology librarian I did try to talk to um /
11		I did try to get more of the local expertise involved um /
12		in helping us scope out you know sites to crawl //
13		But the way she always sort of implemented that was /

Line	Speaker	Utterance
14		she ended up setting us up with some of the very people who /
15		I think these activist groups feel were complicit in what's gone on since then //
16		And one of them I distinctly recall /
17		told me straight up in our meeting that he doesn't think I should be crawling these activist groups
18	Ed	Really?
19	Jack	Because he doesn't find them credible //

98 Articles

Using Gee's making strange tool again to uncover the context of this excerpt draws our attention to the use of the phrases "crawling and selecting" (line 2) and "scope out… sites to crawl" (line 12). "Crawling" refers to the behaviour of software used to collect content from the Web. This software is tradition- ally referred to as a *spider* because it automatically and recursively follows links in web content for an amount of time that is determined by the scope it is given. However, the software needs to be told where to start crawling. Jack sought advice from local domain experts (lines 10–11) in determining where the software should begin crawling and for how long, but he also indicates that he is planning to do more work with the activist communities he is documenting. Jack reflects that activist communities may be concerned with how he has selected content for the archive. He also significantly discusses credibility, in lines 16–19, as a measure of what should (and should not) be in the archive. The web archive is shown to be a contested political space. We also see several relationships being teased out here: the relationship between Jack and other

members of the university community; the relationship between Jack and the community of activists who are working to stop fracking in the state; the relationship between Jack and the software that is performing the archiving activity; and finally, the relationship between Jack and the interviewer (myself) as I orient to his description of how the credibility of the fracking archive had been called into question (line 18). The two communities, of activists and of the university, are presented as being at odds; but elsewhere in the interview, Jack talks about an overlap between them (university members who are also activists like himself). Tracing this network of political agendas and associations is tied up in the work of selecting which websites to archive and is part of the figured world that Jack and I are building in this interview. Doing the work of crawling the Web and appraising web content inscribes these political dynamics into the archive and presents an opportunity to reflect on what they are.

5.3. Discussion

Gee's seven building tasks provided a clarifying lens for studying the discourse that emerged from these interviews. Specifically, Gee's relationships building tool helped trace the connections between archivists, their colleagues, their institutions, and the creators of web content out in the world. These relationships mark pathways of mutual engagement and illuminate how appraisal decisions, or acts of archivalization, are made as part of a community of practice. Carly and Jack's comments about mentoring relationships are of particular interest because they present historical relationships that extended in time, rather than relationships anchored in individual workplaces. Both archivists drew on mentoring experiences when speaking about how they made appraisal decisions. It is interesting that, in both cases, the experiences involved not web archiving but the archiving of physical documents and oral histories. The appraisal decisions were oriented around the purpose or func-

tion of the archives as political agents rather than around the specifics of the Web as a medium. The archivists engaged in translation work to map their experience with archiving print material to help them make determinations about whether to archive web content. The inherent political dimension to these relationships was another feature that emerged from the discourse. Hierarchical relationships within the organizations operated to buttress appraisal or validate appraisal decisions, but we also saw significant relationships between the archivists and the communities or individuals that were being documented. In Carly's case, this relationship was embedded in the professional discourse around government documents work and the operations of government. In the United States, government documents are produced within a legal framework where they are considered part of public discourse and the public domain. In Jim and Jack's cases, there was an awareness of a need for more interaction with the creators of the documents being archived. The role of the archivist in relation to those being documented forms part of a complex terrain that the archivist must navigate in doing web archiving work. The organizational and community relationships intersected with each other to generate productive and destabilizing effects. Developing practices archivists can follow as they go about making these connections with content creators on the Web, or as Jack says, getting more "contact and dialogue" with content creators, is marked as a potential area for further methodological and design work, especially with regard to establishing trust relationships on the Web (Gracy, 2004; Neal, 2002).

The approach of tracing relationships, exhibited here, has much in common with the methods offered by actor network theory (ANT) (Law, 2009), with the important distinction that non-human actors are not part of this discussion. ANT suggests that one way to extend or enrich this work would be to explore how artifacts such as policies, software tools, standards, and services fit into this network of relationships and how they figure into moments of translation in the work of web archiving. In

addition, the concept of a community of practice provides guidance for mapping the interactions and practices of web archiving work. Wenger describes a community of practice as a process of "negotiated meaning" that is achieved through the participation of its members in some joint enterprise. But participation is not the whole story; negotiated meaning is also dependent on something Wenger calls *reification* , which he describes as "the process of giving form to our experience by producing objects that congeal this experience into 'thingness.' In so doing we create points of focus around which the negotiation of meaning becomes organized" (Wenger, 1998, p. 58). Wenger's idea of reification and its role in building communities of practice suggest that consideration of the artifacts of web archiving could enrich this picture of the relationships involved in appraisal work in web archives. This move also nicely parallels Gee's inclusion of non-linguistic elements into the analytical scope provided by CDA. Recall Carly's memorable description of the work she did with her mentor to pile up newspapers and clip articles that held hints or clues about documents they needed to track down for the archive. Activities like this are examples of a shared repertoire that knits together participative and reifying elements of a community of practice. It is important to remember that Carly felt that she lacked an analogue to this practice in her work with web archives. Moving beyond the interview and into participant observation in the context of a case study is one way of exploring this gap.

Considering the political and the material dimensions of appraisal talk suggests that an architectural shift has taken place in the movement from archiving physical media such as documents, photographs, and disks to archiving networked resources such as web pages, websites, and web platforms. Physical media require some form of hand-off, where an archives gains possession of material either through donation or by some other means. This often entails significant work by the archivist, who is often involved in the physical transfer of materials and the negotiation of a deed of gift

that serves as a contract between the archives and the individual or organization that currently owns the material. The architecture of the Web dissolves this traditional relationship because the content can be immediately acquired using the Internet and the Hypertext Transfer Protocol (HTTP). Additionally, web archiving software allows materials to be rapidly collected in bulk, often without any interaction with the publisher or owner of the content. On the Web, the idea that records will be transferred to the archives when they are no longer actively used no longer applies because it is precisely at the moment when records are removed from the Web that they become unavailable to the web archivist, at least with our current set of tools and practices. This revolution in record transfer technologies suggests an inversion of current web archiving practice and a realignment of traditional donor relations, in which web publishers reach out to web archives to have their websites collected prior to them being turned off. The way web archivists talk about their appraisal processes shows that, despite their relative isolation, they work within dynamic and distributed communities of practice that are extensions of a longer trajectory of appraisal in archives. And yet, at the same time, the architecture of the Web and its affordances for access have disrupted the traditional relationship of trust between the donor and the archives. Access to the appropriate tools grants the archivist the ability to easily collect content for the archives with very little interaction with the content owner. This means it is more important than ever to consider the positionality of the archives in relation to the documented entity when deciding what to archive on the Web (Jimerson, 2009). It also suggests that there are opportunities for bridging this gap by becoming participating members of the communities we document, including them in our communities of practice, and developing tools and strategies that help us establish these connections.

> Appraisal brings into sharpest focus the power wielded by archivists the power of what the French philosopher Jacques Derrida calls consigna-

> tion. Which stories will be consigned to the archive and which will not. This power of the storyteller is ultimately a political power. Which is why, in a democracy, society must find ways of holding archivists accountable for their appraisal decisions. (Harris, 1998, pp. 48–50)

As Verne Harris indicates here, accountability is an ever-important dimension to the work of an archives. But engagement in a community of practice that includes content creators as dynamic and complex participants presents challenges for the archivist who works with the Web. What are the pathways of trust in web archives? How do we enact and map them? While these have been perennial challenges for the archival community, they are placed into stark relief in web archives because of the modes of acquisition that often involve the record creator only minimally, if at all. Finally, echoing Emily Maemura's point about the importance of documenting provenance in web archives, we must recognize these moments of archivalization as necessary elements of archival practice on the Web (Maemura et al., 2018). Yet, while provenance looks backwards in time to reconstruct relationships between records and the world, appraisal looks forwards to actively construct them. Appraisal in web archives is charged with an architectural tension, as the Web's access protocols and global namespace collapse expected relations between archivists and records creators. The next step for researchers and archivists working in and with web archives must be to examine how current tools and practices can mend this architectural divide and to establish a social web of trust that determines how a particular set of records ends up in a web archives.

6. Seeing Software

When trying to understand how archivists decide what to collect from the web (appraisal) it is important to get a clear sense of what web archives *are*. But what *are* web archives, really? My review of the research literature and interviews with practitioners thus far have suggested an uncontroversial answer to this question: web archives are the places on the web that collect, preserve and provide access to what other places on the web looked like at a particular time. Examples of such places include sites like the Internet Archive (Lepore, 2015) as well as national libraries and other members of the International Internet Preservation Consortium who assemble a particular set of technologies and practices to crawl, store and "play back" web content (Masanès, 2006b). However in this chapter I argue that this is only one among many shapes that web records can take in archives. Consideration of the full spectrum of web archives is critical to understanding how archival appraisal functions when it comes to the web.

Niels Brügger describes the records of web archives as *reborn digital* in the sense that they were *born digital* somewhere on the web, after which they were collected in a web archive, where they are *born again* (Brügger, 2016). However this conceit is undercut by the recognition that *all* web content is constantly being *born again*. Every HTTP request for a URL made by a web browser is a request for a *representation* of a *resource* to be generated and transmitted again (Fielding, 2000 ; Jacobs & Walsh, 2004). It is this complex and delicate sociotechnical orchestration at the heart of the infrastructure of the web that is the very source of the web's precarity. Indeed, it is this precarity that drives the creation of the spaces called "web archives" in the first place.

So when it comes to the *ontology* of web archives there really aren't any simple answers. Archives can be *on* the web. Archives can be *of* the web. The very

architecture of web itself has archival properties and processes that are inherent to it. In her multi-site ethnography of web archives Jessica Ogden argues that web archives are best characterized as web *sites* (places on the web) that enable a specific set of *claim making* practices:

> 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. (Ogden, 2019, p. 8)

Web archives come to exist in order for a varied set of actors to *make claims* about the world. In this regard web archives are not so very different from more traditional archives in which records serve as evidence of specific activities (Cox, 2000), and where even an antelope can become a record once it is placed in a botanical garden (Briet, 2006). However, as this chapter will describe, the infrastructure of the web both prescribes and proscribes a certain set of claim making practices. Web archiving technologies are assembled to set particular claim making activities in motion. But these technologies participate in a broader network of actors, that can be reconfigured and adapted for other purposes. It is in their uses, disuses and misuses that we will discover what web archives are, and the practices of appraisal by which web archives are assembled and maintained.

In this chapter I explore appraisal practices and this question of the ontology of web archives by analyzing my findings from a year long field study with the National Software Reference Library (NSRL). The NSRL site provided a unique glimpse into one of the world's largest known collections of computer software, which over the past twenty years has transitioned from collecting physical media to collecting

directly from web based software distribution platforms like Steam and the Google Play Store.

6.1. Methodology

To understand the methodology employed in this chapter it is helpful to briefly review my analysis of appraisal in web archives thus far. In Chapter 4 I interviewed archivists and technologists to discover how they they performed their work. I used inductive thematic analysis (Braun & Clarke, 2006) to derive a set of interlocking sociotechnical factors that drive the process of appraisal. In Chapter 5 I used Critical Discourse Analysis (J. P. Gee, 2014) to look closely at participants' use of language in these interviews, to see how their language figured appraisal practices in web archives. Results from this study found that appraisal activities take place in a community of practice (Lave & Wenger, 1991), where identity, politics and the positionality of the archive are embedded in a dense network of personal and organizational relationships.

In short, my analysis of appraisal in web archives so far has been focused on language use, as it has moved from an examination of *what* practitioners talk about (the subject matter) to *how* they talk it (the performance). In this chapter I continue this zooming in process by examining appraisal activities in the context of a specific archival setting, the NSRL. The purpose of zooming in my analysis is to two fold: 1) to test the findings that were obtained previously during my interviews; and 2) to gain further insights into what sociotechnical processes work to shape appraisal practice. Participant observation and ethnography provide a method for understanding how practitioners language and actions are composed *together* at a particular site. The findings gathered at this site help ground and further refine the insights that have been gained in the previous two studies, and will complete the triangulation that forms the methodological backbone of my dissertation.

My field study with the National Software Reference Library was made possible by an existing research partnership between the National Institute for Standards and Technology (NIST) and the University of Maryland (UMD). In the summer of 2018 staff members of the National Software Reference Library (NSRL) approached the Maryland Institute for Technology in the Humanities (MITH) to see if there were researchers interested in working with the NSRL. After some initial discussions the NSRL expressed interest in having me help them with their use of the BagIt standard for digital preservation. As part of these initial discussions I learned that the NSRL was actively engaged in a form of web archiving. So I expressed my interest in conducting a field study at the NSRL as part of my work there. The NSRL agreed to this and the research partnership was initiated in August 2018.

My field study lasted for 16 months. During that time I was able to work at NIST's Gaithersburg campus for approximately one day per week. Being physically on site allowed me to attend weekly staff meetings, share office space with NSRL team members, participate in collaborative workspaces such as Slack and Bugzilla, and to generally soak in what it meant to work on the NSRL project, and be a NIST employee. During this time I actively created field notes that documented my own activities and those that I observed. Jottings and photographs taken during my visit were used as source material for longer reflections on how work in the NSRL was being performed (Emerson et al., 2011). As I will discuss in more detail below, after twenty years of activity the NSRL's work extends out into other departments in the Information Technology Lab (ITL) at NIST, and outside of NIST into digital forensics research, law enforcement and intelligence communities. Given this complexity it was essential for my fieldnotes to reflect on not only on what I observed, but also on my own thoughts and feelings as I learned more about the people and the work of the NSRL.

While these jottings and field notes formed the bulk of my research data, I also conducted a series of unstructured interviews (N=12) with NSRL and NIST employees, as well as a few individuals from outside of NIST whose work touched on the activities of the NSRL. These interviews provided me with an opportunity to follow up on salient things that I had learned during participant observation, and field note taking. An informed consent process provided by both a UMD and NIST Institutional Review Board (IRB) gave participants an idea of how their interviews would be used as part of the research, and offered me a singular opportunity to learn more about how NIST employees came to work on the NSRL. As part of that IRB process psuedonyms have been used to identify participants, and have been given the opportunity to qualify or redact any quotations that I have used.

The final category of data that I analyzed was documents related to the NSRL's work. On my first day on the NIST campus one NIST employee remarked to me that "we have no secrets here". Indeed, the NSRL's mission is to serve as a public resource for the law enforcement and digital forensics communities. Many of the NSRL's staff have authored research articles, given conference presentations and been the subject of interviews with the media. Over the last 20 years NIST itself as part of the US Department of Commerce has published a substantial amount of material about the NSRL in government documents and on its website. In the case of information that is already easily identified because of its published nature I have made no attempt to obscure its origins and cited it as any other bibliographic resource.

My analysis of these materials followed a process of transcription and line by line coding using Grounded Theory (Charmaz, 2001). A set of 108 initial codes was inductively generated during the first phase of coding. A log of observations detailing why and when codes were introduced along with general observations on the process was kept. The transcriptions, codes and my log together provided the mate-

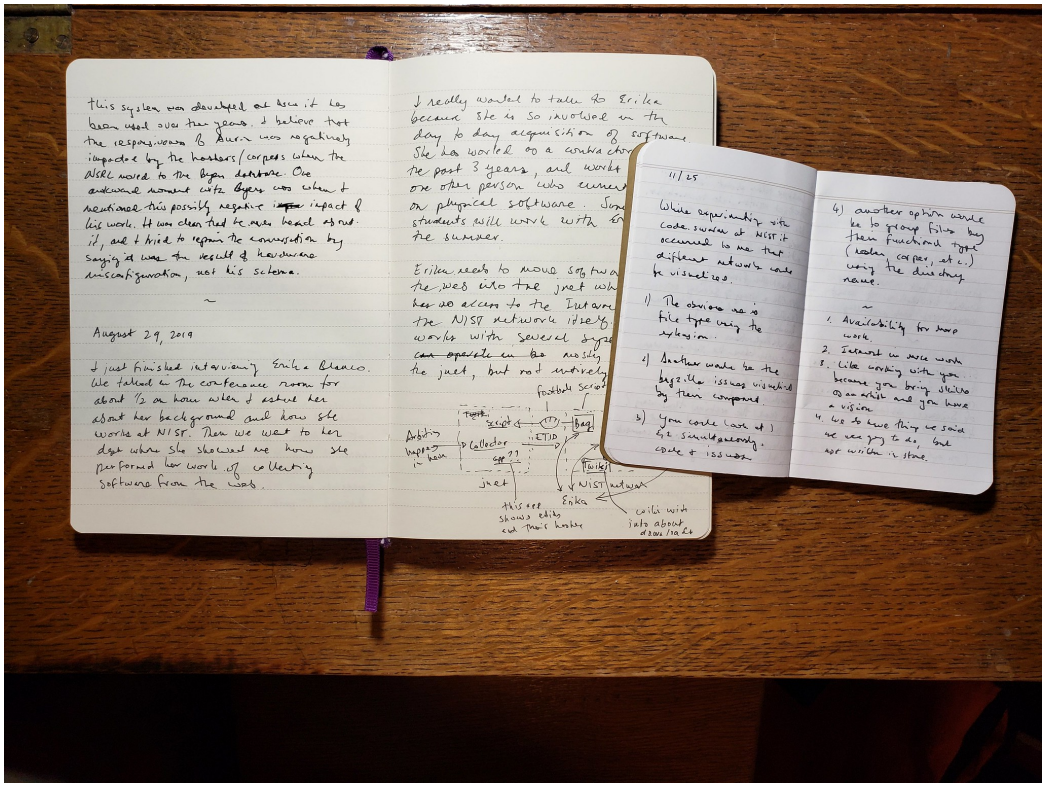

Figure 3: Field Notes and Jottings

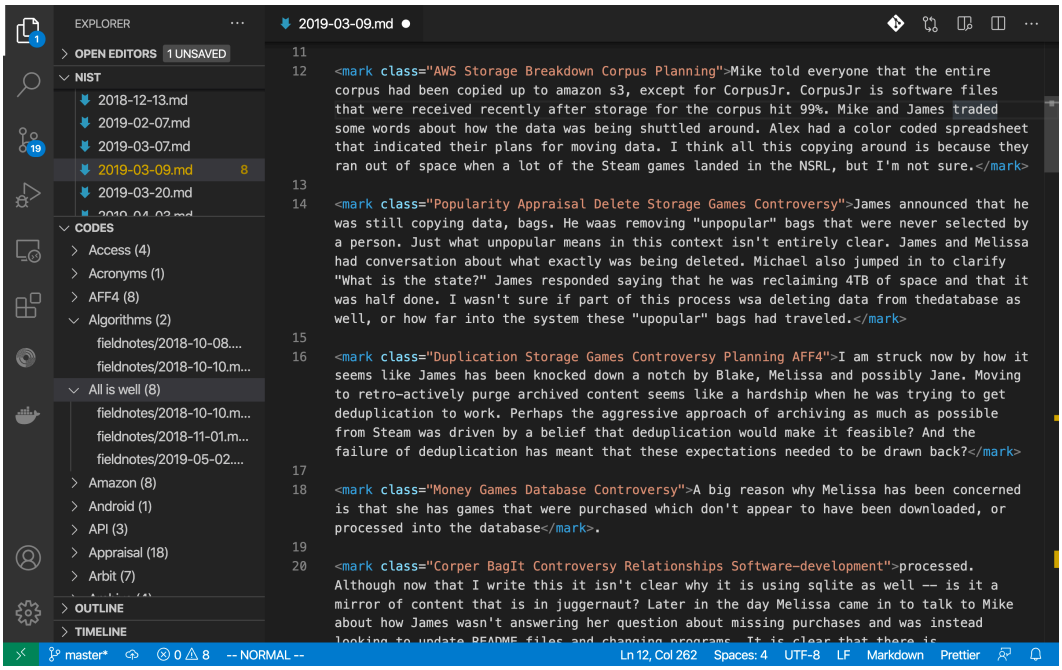

Figure 4: Coding with Anselm

rial for a secondary memoing process for connecting and aligning codes, in order to develop progressively more abstract categories that synthesized concepts for understanding the data. These categories are detailed below in my findings, where they are discussed in the context of selected vignettes.

Classical Grounded Theory required researchers to start with a blank theoretical slate from which to inductively generate new codes, new categories and (ultimately) new theory, directly from their data. This approach was intended to counteract the potential for confirmation bias that gets introduced by analyzing data using pre-established categories or conceptual schemas. If only it were that easy to willfully purge our brains of unwanted prior knowledge and memories. Such a selective amnesia recalls the characters Joel and Clementine in the film Eternal Sunshine of the Spotless Mind, who attempt to erase their memories of each other after a failed relationship, only to find them resurface later in memory fragments, and the actions of those around them. More recent Grounded Theory research continues to emphasize inductive analysis, and to minimize the commitments to preconceived theories, but while recognizing that analysis is a continuous thread that runs throughout the life-cycle of a research project: in the research questions that are asked, in the sites that are chosen, in the data that is gathered, and in the analysis of the collected data itself (Charmaz, 2001 ; Emerson et al., 2011, p. 172). So before diving into some ifs findings it is useful to briefly acknowledge the theoretical commitments that guided the conception of this study.

The principle theoretical commitment taken here is to a sociotechnical approach that understands work (such as archival appraisal) to be mutually constituted, or co-produced, by the interaction between society and technology (Jasanoff, 2006). Sociotechnical theory commits to explanations that compose social and technical factors together, without reducing a problem space using social constructivism or

technological determinism. A particularly important guide is the principle of *symmetry* from Actor-Network Theory (ANT) which encourages researchers to "follow the actors" where actors are any entity (human or non-human) that *makes a difference* in the world. For Latour, understanding and unpacking the dense network of relations between *actants* (a word that does not privilege human or non-humans) is critically important.

> Action is not done under the full control of consciousness; action should rather be felt as a node, a knot, and a conglomerate of many surprising sets of agencies that have to be slowly disentangled. It is this venerable source of uncertainty that we wish to render vivid again in the odd expression of actor-network. (Latour, 2005, p. 44)

Another commitment that was made plain in Chapter 2 is that one useful way of synthesizing the divergent research literature of archival appraisal is through Foucault's idea of *governmentality*. For Foucault, the study of governmentality is not simply concerned with the functions of government, but with knowledge practices, or modes of rationality, that exercise, perform and enact power relations. Archives are key components in a network of sociotechnical memory practices that enable very specific social and political relations (Schwartz & Cook, 2002). So in addition to diversifying the number of actors my study also recognizes the central importance of tracing these power relations, wherever they may lead. Given the central role of that computation plays in web archives it particularly important o understanding these power relations from the perspectives of Critical Data Studies (boyd & Crawford, 2012) and Critical Algorithm Studies (Seaver, 2013). These two theoretical perspectives employ critical theory (e.g. the study of governmentality) to investigate the social conditions of computation when it comes to rest as data, and when it is operationalized in processing (algorithms).

Finally, my analysis has been greatly informed by the significance of narrative and story. These are the stories that my participants shared with me during my fieldwork: in staff meetings, in the midst of doing their work, as we met in the hallway, in notes left in various types of documentation, and in the scheduled interview conversations that they so generously shared with me. In her book *Working the Past* Charlotte Linde develops a framework for understanding how groups of people remember together using stories, and how this storytelling and retelling of stories enacts their present, and projects their futures (Linde, 2009). Attending to the occasions for stories and their telling, such as the times, places and objects involved, helps researchers gain insights into how people understand themselves, their work, and each other. As I will describe below, sometimes the same story is told and retold by multiple participants. Sites for storytelling can be regularized such as weekly meetings or conferences, or unscheduled conversations walking down the hallway, in emails or in Slack channels, in the logs of of version control systems, and issue ticketing systems. In his description of tactics for studying algorithmic systems Nick Seaver encourages researchers to be ethnographic scavengers, and to attend to the many occasions for empirical data gathering when studying algorithms *as* culture (Seaver, 2017). Seaver indicates that given the way algorithms are often guarded, and knowledge of them can be distributed across teams and organizations, it may be the only way to study algorithms as sociotechnical systems.

One final meta note on stories before I get on with mine. It is through its dependence on fieldnotes, and the generation of textual accounts that ethnography as a methodology deeply engages with phenomena in order to elicit understanding. Jottings record brief observations from the field, which get incorporated into fuller notes, which get analyzed, categorized and further refined to produce the textual accounts I will share below. Widening the arc of this meta analysis further it is clear that archives, and web archives, the subject of my study, choose to tell particular stories,

while simultaneously choosing not to tell others. What follows are the stories of appraisal as it happens in a web archive.

6.2. Findings

In each section below I will share a short description of activity or vignette drawn from my time working in the NSRL. These accounts are highlight a particular aspect of feature of archival appraisal in this archive. Each section will be introduced with a few words about this significance before diving into the specific details. The threads of concern that tie these findings together will be outlined in the discussion section that immediately follows.

6.2.1. NIST

Before eliciting some of the activities of the NSRL and their relevance for understanding archival appraisal in web archives it is important to set the stage, and describe a bit more fully what the NSRL is as well as how and when it came to be. The history of the NSRL extends over multiple decades and is an effort by an institution that is over a century old, so this description makes no claim to being complete. The purpose is to adequately ground subsequent descriptions of NSRL activities that are relevant to the study of archival appraisal.

The National Software Reference Library (NSRL) is a project based at the National Institute for Standards and Technology in Gaithersburg, Maryland. For the last 20 years the NSRL has gradually assembled one the largest known collections of computer software in the world (NIST, 2018). In 2000 the NSRL began by collecting "shrink-wrapped" software (physical disks, CDs and DVDs) and storing them in a physical library on NIST's Gaithersburg campus. Digital copies of the media were are also created to serve as a backup and also as the input data for a set of services that the NSRL provides. More recently the NSRL has transitioned to collecting soft-

ware directly from the web as it has became the predominant distribution mechanism for computer software. The NSRL presents a rich site for observing what archiving practices look like in the context of a longer trajectory of archiving physical media.

While the NSRL has been housed at NIST since its creation in 2000, it was initially created with funding by a diverse set of federal government agencies that found common interest in collecting computer software for law enforcement purposes. This excerpt from a NIST annual report that was published shortly after the NSRL's inception describes these initial actors and their motivations:

> Computer forensics is rapidly becoming recognized by the legal and law enforcement communities as a science on a par with the other forensic sciences. As this trend continues, it will become even more important to handle and examine computer evidence properly. The National Institute of Justice, the Federal Bureau of Investigation, and the Department of Defense Computer Forensics Laboratory asked [the Information Technology Laboratory] ITL to provide a neutral and technically proficient source of reference data and test procedures. (NIST, 2001, p. 10)

The dataset mentioned here is *NIST Special Database 28* otherwise known as the *National Software Reference Library Reference Data Set* (RDS) which is described on the NIST website as "an example of the application of technology to investigate crimes involving computers, such as child pornography, racketeering, cyber-attacks, illegal gambling, Internet fraud, and software piracy" (NIST, 2020b). The NSRL is the successor to an FBI project called the Known File Filter (KFF) which provided an inventory of digital signatures of software files. The FBI contacted NIST in 2000 to ask them to collect software and generate file fixity metadata about the files contained by the software. In the words of Doug White, the current NSRL Project

Lead, NIST was chosen as a site for the NSRL because it would make it (the RDS) "from an unbiased source, because we are not law enforcement, and we're not a vendor. It's collected in a transparent manner. We share everything that we do, as far as procedures. It's collected in a scientific manner. It's collected in a timely manner. And it's easily used in digital forensics tools." (Lyle, 2017) While work on the NSRL was initiated just prior to September 11 terrorist attacks, it

> Since the tragedy of September 11, 2001, ITL's work in computer forensics has taken on added significance and impetus. Sound computer forensics practices are a key to finding and delivering court-permissible evidence when computers are used in the commission of a crime. Our program has two components: the National Software Reference Library (NSRL) and Computer Forensics Tool Testing (CFTT). [NIST (2002b); p. 12]

The numbering of Special Database #28 clearly identifies it as part of a series of "databases" that NIST publishes in its Standard Reference Data Catalog. In 1968 the Standard Reference Data Act granted NIST the ability to collect and sell collections of reference data that are "quantitative information, related to a measurable physical or chemical property of a substance or system of substances of known composition and structure". In 2017 this definition was updated by the American Innovation and Competitiveness Act, which broadened the definition of SRD to include measurements involving *digital* objects: "1 or more digital data objects that serve to calibrate or characterize the performance of a detection or measurement system".

While it was initially distributed on CD to subscribers, since March 2018 Special Database 28 is only available as a download from their website. NIST is able to provide this access because, unlike a traditional library or archive, they do not make the contents of the NSRL directly available to its users. Instead they process the

collected software as data to create the *NSRL Reference Data Set* (RDS), which is a collection of digital signatures, or *hashes*, of all the files, from all the software that NIST has collected since work on the NSRL began. These digital signatures are then used to identify files of interest during forensics investigations. This processing will be covered in more detail below, but the motivating principle for the NSRL is to provide reference data for forensics tools that allow their operators to rapidly inspect computer file systems looking for unique files. Figure 3 is an illustration of the use of the RDS taken from a presentation made by the NSRL at the Regional Computer Forensics Group at George Mason University shortly after the creation of the NSRL (Lyle, 2002). While looking somewhat dated, the same image was used 16 years later to describe the NSRL's efforts to collect video games. As of June 2020, the publicly released NSRL RDS tracked over 250 million hashes for files that were part of 213,770 versions of 171,567 software packages, from approximately 70,000 publishers.

6.2.2. The Scene

Understanding sociotechnical systems requires that researchers avoid privileging purely social or technical explanations. Phenomena are to be understood instead as assemblages, networks or knots of activity that include a heterogeneous set of actors. One way of achieving this admittedly abstract goal is to *multiply* the number sources of action that are attended to, by granting agency not only to humans but also to "non-human" things or materials. For Latour anything that creates a *difference* for another agent in the world needs to be examined when studying "social" phenomena (Latour, 2005). Simply denoting some field activity as social is not sufficient–it needs to be traced and related to other actors. Social explanations must not be restricted to analyzing only human agents because doing so renders any analysis circular: X is a social phenomena because there are people doing this set of things, and humans are

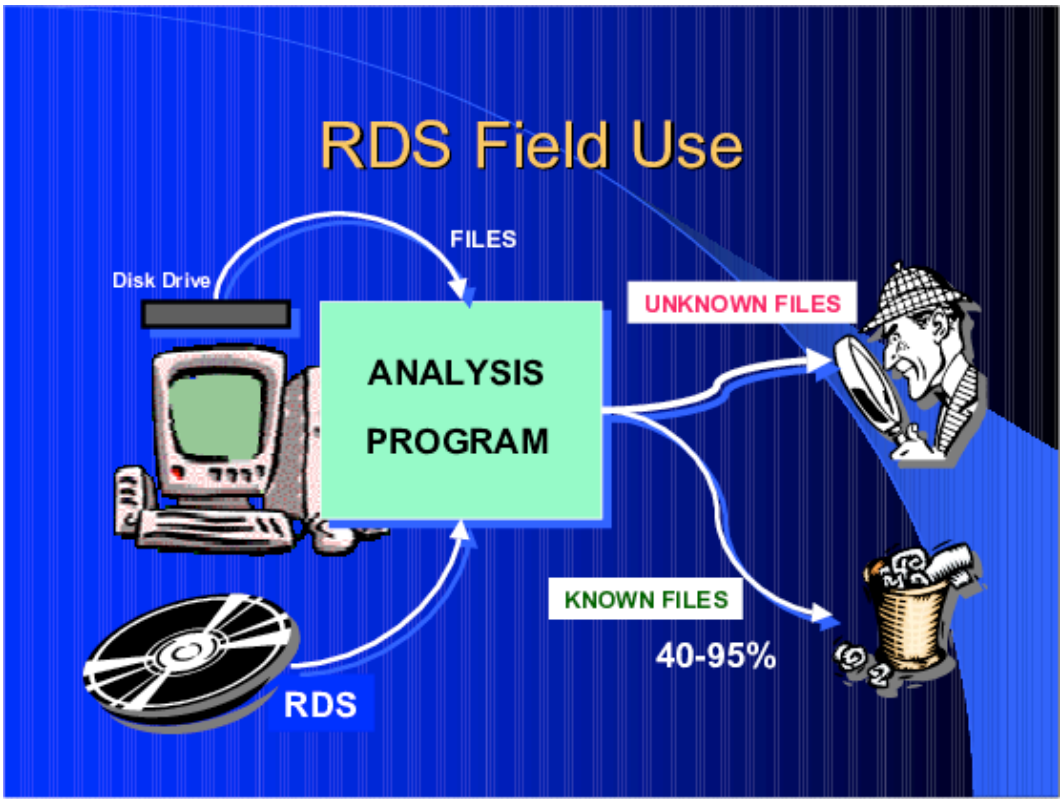

Figure 5: RDS Field Use

part of the social world, ergo X is a social phenomena.

Multiplying the types of agents that are allowed to participate in a network of activity paints a more dynamic picture of the social, and the networks of activity that constitute the social. This next section picks up where the previous section left off to describe a few of the agents that are involved in the creation of the NSRL Reference Data Set. Naming these agents and drawing their network of relations in the setting of the NSRL is important for understanding the shapes that appraisal will take.

I often chose to work at NIST on Thursdays because it was the day scheduled for the NSRL weekly staff meeting. I would take the Red line Metro from where I live in Silver Spring, down through Washington DC, and out again to the Shady Grove station, where the NIST shuttle would pick me up along with other commuting employees, and drive for about 15 minutes along Route 370 to Route 270, and then onto the NIST campus. At the gates to NIST an armed security guard would enter the shuttle bus and individually check everyone's staff ID. If you didn't have an ID you would be ejected from the shuttle. After moving through the get the shuttle would circle along the north eastern portion of the 579 acre campus, and stop at Building 101, the Administration Building (Figure 4). I then had a ten minute walk south, to the Chemistry Building 222 (Figure 5), where the NSRL team had office space. It never stopped feeling incongruous to walk into the Chemistry Building because it was clear from the research posters I saw as I walked by people's offices that people were interested in things like network security, cyberinfrastructure, machine learning and quantum computing. Presumably the work on Chemistry continued on in another building. It felt as if the activities going on inside this building had exceeded or overflowed its name. No doubt the post World War 2, brutalist architecture contributed to this sense of time slippage. The entire NIST campus was designated a "Historic District" in 2016, and it clearly felt as if the buildings were frozen in a

specific historical moment, like a museum piece.

Figure 6: Administration Building 101

The staff meetings were usually attended by six to eight people, although there were a few exceptions to this which will be discussed in sections below. Attendees were usually NIST staff who worked directly on the NSRL project. But not all the participants were from the same office within NIST. The NSRL project is housed within the Software Quality Group, and most, but not all, the staff at the weekly meeting were from there. For example James, one of my primary contacts on the project, was from the Computer Security Division, which, like the Software Quality Group, was part of NIST's sprawling Information Technology Lab (ITL), but not part of the Software and Systems Division, it was part of the Computer Security Division. The Chemistry Building was rectangular, and had three floors filled on its perimeter with small offices, each having one or two desks in them. The core of the building filled with group office spaces and meeting rooms, and it was one of these rooms in

Figure 7: Chemistry Building 222

the core of the building where the weekly staff meeting was held. It is hard to imagine a building that embodied the essence of bureaucracy more, in its complexity, and quite literally the power of desks or, as I would come to experience, computer desktops (Figures 6 and 7).

Figures 6 and 7: Chemistry Building Hallways

If the names on the buildings and the many levels of administrative hierarchy weren't confusing enough, the staff meetings immediately immersed me in a tumult of unfamiliar names and language. Several of the staff have worked at NIST, and on the NSRL project itself, for decades. A number of them came to the project as student interns, and stayed on afterwards after finishing undergraduate or graduate studies to work as full time staff. Some employees grew up in families where parent worked at NIST, and had siblings that worked elsewhere at NIST. Needless to say, the NSRL staff had lived and breathed the NIST culture and had its activities interwoven with their lives. Even after a year, each meeting was an opportunity for me to learn some new name for a system, process, department or person. Unless otherwise noted all names for individuals at NIST are psuedonyms.

All the individuals named in the vignettes below are pseudonyms, and they will be introduced as needed. But to understand what these people are doing it is helpful to have a picture of the types of processing that goes on inside the NSRL. As discussed above, the public facing output of the NSRL is the Reference Data Set (RDS), which

is the published set of file hashes and their metadata for software that has been collected in the NSRL. The staff meeting was almost always concerned with updates about the performance of the processing pipeline, and especially with the details of its deployment and maintenance.

Since 2001 the NSRL staff has developed and actively used its own set of bespoke software utilities, databases and coordinated workflows for generating the RDS on a quarterly basis. These utilities had distinct names, but their functionalities were fluid, in that they were constantly adapted, by fixing bugs, or adjusting them to work on new types of data. I came to understand these components only partially during my time at NIST, while listening to the staff talk about and perform their work. The NSRL was actively engaged with collecting software from gaming platforms, so many of the pieces of the puzzle that I learned about related to them. These software actors operated semi-autonomously, and sometimes when their human counterparts could not, as during the 35 day shutdown of the federal government. Here is a listing, or glossary, of some of these software agents:

Juggernaut: The database inventory of all the software that has been acquired by the NSRL over its 20 year history. The database has seen four generations of development, the last of which was initiated just prior to my year with the NSRL. Juggernaut is a PostgreSQL database that records metadata about the software that has been collected (name, version, manufacturer, operating system) as well as metadata about the files themselves (path and hash value). All of the software applications talk to Juggernaut either directly or indirectly. It is named after the X-Men comic book character.

Gateways: Gateways are units of code and machines that are used to collect software from various platforms such as Steam, Blizzard, Origin, Google Play and the Apple Store. Since each platform makes its data available differently, and the NSRL

has varying needs of them, custom Gateway Scripts or sometimes Applications are written for each, which allow software to be downloaded and packaged up for delivery into the NSRL. These downloads often need to happen over a separate Comcast connection to the Internet since NIST's own network infrastructure blocks particular gaming platforms and software distribution sites.

Bags: These are packages of files that constitute a unit of downloaded software that has been assembled by one of the *Gateways*. Each "bag" is a zip file with a file path structure specified by BagIt (Kunze et al., 2018) which includes both the software files that were downloaded as well as metadata specific to the platform, such as the Steam Application ID (STAID).

JNet: The majority of NSRL's infrastructure runs on servers that are closed off from the Internet and from NIST's own network (INet). This network is known as the JNet in homage to the INet (NIST's network). The JNet is also a place, a set of rooms where machines and people are colocated so that they can work together on the work of the NSRL. Some people have offices in other spaces, but also use desks or cubes in the JNet room.

Football: A physical storage device that is used to move the *Bags* created by *Gateways* from the machine connected to the Comcast line to the Internet, into the JNet.

Replicators: A daemon that runs on *Gateway* computers where the *Football* lands, which copies the *Bags* into the appropriate network attached storage location where a *Corper* can pick it up.

Corpers: A distributed set of processes that look at *Bags* (or previously *Images*) and copies their payload data into *The Corpus*, which is a collection of all the software files ever discovered by the NSRL stored on Network Attached Storage. The Corper "cracks open" the initial package of data by mounting it, and extracting the files that are contained within.

Unpackers: A distributed set of processes that look for a wide variety of container file formats such as ZIP or JAR, and unpacks their files to the *Corpus*. This process is done recursively since any files that are unpacked could also be containers for other files. Metadata for new files are written to *Juggernaut* as well as to *The Corpus*.

Hashers: These are processes which look for files that have been unpacked but have not yet had their checksums (MD5, SHA1, CRC) calculated. Once a file has been "hashed" the checksum is stored in *Juggernaut* as well as *The Corpus*.

Imagers: Imagers are workstations that are used to create *Images* which image physical media (disks, CDs, etc) or snapshot virtual machine file systems in order to capture the state of some software after it has been installed. This can be common for operating system software like Microsoft Windows where running an installer triggers the download of additional data from the Internet.

Diskprint: A disk image of a system that has had software installed and then run in a controlled manner in order to collect additional files from the network.

Arbiter: A quality assurance system and process called *Arbiting* for deciding whether collected software, and its metadata match, and are ready to be released as part of an RDS.

Collector: An older system for collecting information about physical media. A critical function that this performs is to create an *Evidence Tracking ID* (ETID) that uniquely identifies the physical item.

Repositories: A set of file metadata and files that is roughly grouped according to its source: iRepo (the original incoming storage), bRepo (bagged gaming objects), mRepo (mobile objects), dRepo (). These "repos" are Git repositories repositories, which became easier to manage by being broken apart.

RDS: A public release of the NSRL's Reference Data Set, which is made available on the NIST website for download as a set of ISO 9660 images, as well as some additional metadata and zip files that minimize some of the data to ease use. Staff often talk about their work using the numerical version of the release instead of saying "the RDS". So for example they will say "two six four" instead of RDS version 2.64.

Library: A physical room in the core of Building 222 which houses the collection of physical media on collapsible shelves. The items are filed according to their *Evidence Tracking Identifier* (ETID), that is accession order, or the order in which the software was acquired.

In addition to their bespoke software the NSRL actively use some other services as part of their work.

Slack: A NIST wide collaborative work environment. There is a specific workspace for the NSRL where staff ask questions, share information and have conversations.

Twiki: A wiki environment which is used to document aspects of the NSRL's work.

Bugzilla: A ticketing system that is used by the NSRL to track tasks that need to be performed. These tasks can include actions to download software, maintenance of infrastructure, and the fixing bugs in existing code and database schema.

GitLab: A NIST wide environment for sharing code and data using the Git version control software. It runs on NIST's network infrastructure and is only available to NIST employees. Some contents from the GitLab are mirrored to their public *GitHub* organization account.

It bears repeating that my understanding of these architectural pieces evolved over time as I was at the NSRL, and never fully settled. Partly this was the result of the architecture being a moving target as it was updated and changed over the year,

and had been grown over the two decades of work to accomodate multiple software delivery mechanisms, from disks to the web. Even while writing this account here I found myself emailing staff for clarification about certain pieces such as the purpose and contents of the various *Repositories* and its relationship to the *Corpus*. You can see this change in my understanding in these sketches from my fieldnotes as I attempted to sketch out what I knew of the architecture.

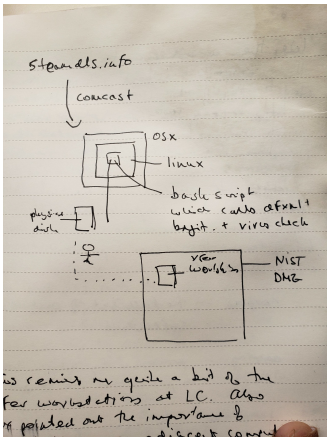

Figure 8: Architecture Sketch: October 10, 2018

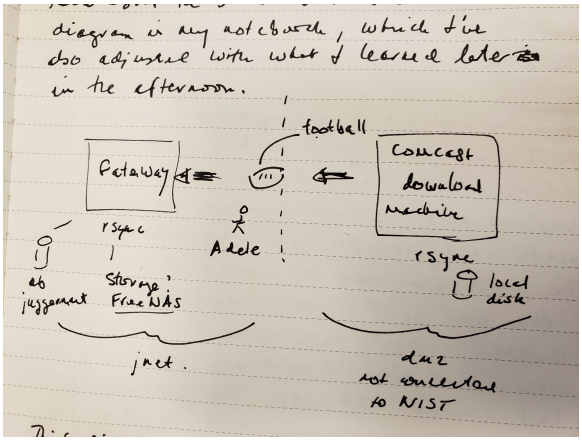

Figure 9: Architecture Sketch: December 13, 2018

6.2.3. Google Play

Over the last 10 years the NSRL has increasingly moved from collecting software from physical media (disks, CDs) to collecting software from the network by down-

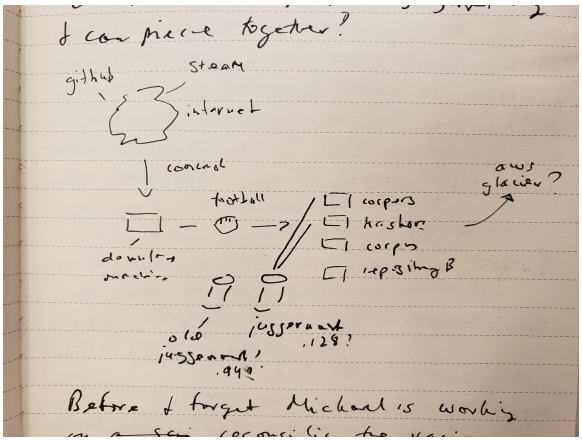

Figure 10: Architecture Sketch: March 3, 2019

loading software directly from the web. Over the same time period more and more software applications have been developed for, and deployed to, mobile operating systems such as Android and Apple iOS. To adapt to this new software environment the NSRL has developed a variety of tools and practices for *downloading* software from the web.

The following segment draws on an interview with Darius, who is a researcher and developer who helped the NSRL enhance its ability to collect Android mobile applications. Darius' prior understanding of how mobile applications are developed and distributed intersects with the directive he was given to collect "popular apps". The identity of the collected software, and its particular digital manifestation, is a key factor in evaluating how to perform the collection. This segment highlights the role that expertise and situated knowledge (Haraway, 1988) play in collecting from the web. It also demonstrates how this collecting activity is mediated by bespoke software that encodes and performs a particular interpretation of appraisal.

Darius started in the NSRL roughly at the same time I did in 2018. He came from France as a Guest Researcher with TELECOM Nancy, which is a public school of engineering that is part of Université de Lorraine. There are many guest researchers at NIST, and incidentally I was one too. By the time I completed by field study at the

NSRL Darius had already returned to work in Paris. The NSRL funding for Guest Researchers had dried up. After graduating from TELECOM Nancy Darius had worked briefly at a startup and then for Oracle Consulting. He told me that at the time he missed doing "research" work, which was having the time to develop software properly, without the constraints and pressures of business development. He learned from some classmates that there were opportunities for TELECOM Nancy graduates to work at NIST, and so he applied. It was not uncommon for me to hear French being spoken in the hallways, and in offices as I walked by, so I knew that Darius was one of several French nationals who had found a way to working at NIST, and specifically within the Software Quality Group.

On joining the NSRL Darius knew that he would be working to help collect mobile applications or "apps". He had significant experience developing and distributing Android apps prior to coming to the NSRL; but he did not have any experience working in the field of digital forensics. His practical experience with how apps were built and made available was critical to his work with the NSRL. He was instructed by NSRL management to focus on the Google Play Store in order to collect "popular" Android apps. He spent a significant amount of time researching how to download apps from Google Play. This involved "testing" the various mechanisms, and comparing the resulting downloads:

> So first I have to make tests. For example: is the application the same if we are downloading it from NIST, from a phone, from different phones, or from different accounts? I had to make a lot of test cases in order to see if the integrity of the app was respected–if it was always the same. So it took two months because we have to think about all the scenarios. So we had to switch accounts, switch the operating system, switch from mobile to tablet to a watch. The result of all this was that the ap-

> plications are still the same… My main objective was to download a lot of apps and the easiest way to download a lot of apps is directly on a Linux computer and not with an Android phone or a lot of Android phones and you have to click it's tedious.

Darius' experimentation led him to develop his own software for downloading apps from the Google Play Store API which combined an approach of "web scraping", directly using the Google Play Store API, as well using the Android Asset Packaging Tool (AAPT) to extracting software and metadata from the downloaded Android Package (APK) file. During his previous work as an Android app developer Darius had acquired some expertise building APK files for distribution on the Google Play Store. Critical considerations that his software application needed to address were needing to authenticate as actual user, which country and region to use, how many concurrent downloads to perform, how rapidly to download the applications. These last two issues were especially important because Google would block a downloading client if it appeared to be working in parallel (multiple downloads at a time) or working to a regular schedule. Darius needed to introduce and tune a random wait time between downloads so that it would not get blocked.

Another significant consideration when building this download tool was how it should interpret the idea of *popularity*. Darius told me he was instructed to download the most popular applications because these are the applications that would be most likely to appear in forensics investigations. The Google Play Store website includes a general list of the most popular applications, and then lists the popular apps in a set of categories such as Business, Comics, Casino, Music, Sports, etc. Darius surmised that the popularity ranking was based on the number of downloads–but how the ranking worked internally was unknown. Similarly the nature of the categories, whether they overlapped, and how many applications were contained in

them, was not clear. The rate at which the popularity rankings changed was also not known. But Darius was aware that some apps like the Facebook app could be updated multiple times per day. Darius designed his utility to bulk download apps by category, by collecting application ids and their metadata from the Google Play Store website, and then using the API and AAPT utility to obtain the APK and extract information from it. The utility had a user interface designed for NSRL staff who could select a category and let it run over night. It typically could collect about 10,000 APK files in 10 hours, where each APK was between 20 and 100 megabytes (MB).

Darius encountered major difficulty downloading "paid apps". These are software applications that required payment to the Google Play Store.

> Actually the process for the paid apps is tricky because we are using an emulator to emulate our accounts. The account on a phone and the account the app uses are connected. There are a lot of processes between Melissa and the NIST administration to give us money. So the account is connected with a NIST bank account. For example we put $50 in the Google Play account and then Melissa uses that account to download an app by hand. Then I'm able to emulate that phone and then access to the store part of the phone using the desktop because the emulator and the desktop are sharing a file system together. And then I get the application, the APK. So it's kind of tricky but it's the only way we found.

So while the process is quite streamlined for bulk processing of free apps it is more complicated for paid apps which require the app to be purchased. Purchasing requires a series of fiscal transactions from NIST's adminstrators that must be done more deliberately. Darius mentioned that this process was very slow, but because

there was not a big budget for purchasing Android apps yet it was not a problem.

6.2.4. Game Platforms

As Darius' experience illustrates above, even with a clear mandate to collect popular software, there is a marked tension between collecting "free" versus "paid" Android applications. In addition to collecting Android apps the NSRL staff is actively engaged in collecting *video games* from several web platforms such as Steam, Origin, Epic and Blizzard. These platforms require similar approaches to those taken by Darius however, unlike Android apps, there is significant pressure to collect *popular* games which cost money. The introduction of payments requires creative administrative problem solving, and new requirements for *traceability* which conflict with automated bulk downloading. These games are distributed using the web, but the game websites themselves are complex sociotechnical *platforms* with their own particular modes of distribution, moderation and governance (Gillespie, 2010 , 2018).

Staff meetings are often a time when Melissa and Vlad coordinate the purchase of games. As the "librarian" Melissa is responsible for the management of the library collection, especially when it comes to the acquisition of specific software. Vlad is a technician who works with Melissa to perform the purchase and collection of games. Melissa has been at the NSRL for ten years since she worked as an intern doing metadata entry during high school. Vlad also began work as a student, but in the Computer Security Division, where he did data entry work for the National Vulnerabilities Database (NVD). He moved over to work full time for NVD on graduating and then was recruited to the NSRL because of his experience with video games. Both Melissa and Vlad, as well as several other members of the NSRL team, collect and play video games recreationally outside of NIST. Vlad's experience downloading games and running Steam game servers, and how to build out "gaming systems" for playing games were instrumental to NSRL's game collecting.

When it was her turn to give an update Melissa announced "So, I went to Safeway this morning and got some gift cards. I'll send over the codes to you Vlad." Vlad replied "Awesome, there's a sale coming up next week." Other team members smiled and chuckled as Melissa described how it felt a bit awkward when checking out to be purchasing such a large stack of Steam gift cards. These gift cards were needed to purchase games on the Steam platform since the NSRL's purchasing credit card (P-Card) blocked purchases at Steam. Blocking Steam puchases was a NIST wide policy, which couldn't be disabled for a specific unit like the NSRL. Since Steam purchasing was a such high priority Melissa worked with the NIST administration in order to get approval to use the NIST P-Card to buy up to $3,500.00 gift cards a month for use in purchasing Steam games. Part of this agreement was that the purchases be traceable, in case of an audit.

Figure 11: Steam Gift Cards

Ordinarily Melissa would purchase a specific piece of software online with the P-Card, and would then create a Bugzilla issue for Victoria to download the game using a license code that she would email to her separately. This issue contained a *Requisition Number* which could be used in combination with the Evidence Tracking ID (created by Victoria in NSRL's *Collector* application) to *trace* the order history.

In the case of Steam there was no direct order history for a purchased game since

the item that had been purchased was the gift card, and the gift card was then used to acquire the game. To account for this level of indirection Melissa maintained a spreadsheet that listed the Gift Card numbers that had been purchased, which had a column for an Evidence Tracking ID (ETID). Vlad would use lists of popular games to select new games, and when he purchased a game with a gift card code he would email Melissa the Steam Application ID (STAID) for the purchased game and the Gift Card number that he used. Melissa would then need to watch a daily custom report that Mike (an NSRL software developer) created which directly queried *Juggernaut* (the database) to list any new STAIDs that had been automatically downloaded, unpacked, corped and hashed, with its ETID. Melissa would then add the ETID to her spreadsheet. This entire process could take weeks to complete since the Steam workflow was only partially automated, and did not run on an explicit schedule.

In a meeting a few months later James (a software developer and researcher from the Computer Security Division who also worked in the NSRL) somewhat abruptly announced that he was in the middle of deleting "unpopular" bags. Melissa asked James specifically which bags were being deleted, and before he could answer Blake (another long time software developer for the NSRL from the Cybersecurity and Privacy Applications Division) asked "Yes, what is the state of these?". In the ensuing conversation it became clear that after the holiday break, and the 35 day shutdown of the Federal Government, the NSRL had run out of storage space. Part of the reason for this was that the software process for downloading Steam games included logic for downloading popular and unpopular games in addition to ones that had been specifically selected by Melissa and Vlad. James had been asked to turn off the automated downloading of Steam games, and to delete the "unpopular" apps that had been downloaded, which would free up 4 TB of storage space.

A few years prior in 2017, when the Steam work was just getting underway, Melissa asked for a *Juggernaut* table that would track the Steam Application ID (STAID) for the game and "some sort of Steam receipt". This table was never realized, however a table was created for recording the STAID along with whether the game was popular.

```
CREATE TABLE "humans"."STEAM_APP"(
  "steam_app_id" Serial NOT NULL,
  "popular" Boolean DEFAULT false NOT NULL,
  "purchased" Boolean DEFAULT false NOT NULL,
  "freepromo" Boolean DEFAULT false NOT NULL
);
```

Figure 12: The STEAM_APP Table (SQL)

This SQL to create the table *STEAM_APP* designates the *popular* relation as a *Boolean* and *NOT NULL*. This means that popularity for every Steam game is either True (popular) or False (unpopular). Later when James was needing ot delete unpopular games to conserve space in NSRL's storage he proposed that the *NOT NULL* constraint be relaxed so that Steam games could be designated popular, unpopular or unknown. This was because new games sometimes lacked statistics around popularlity. This prompted Mike, one of the NSRL's lead developers to remark in a Bugzilla ticket:

> I don't think a trichotomy is needed. I think it's too much work to try to make the automated classification of "popular" perfect, for items that we collect quarterly. If an item isn't rated "popular" in the 24 hours before the script runs, no big deal, if it becomes popular in the next 3 months, we'll get it that quarter instead of this quarter. Using 5000 simultaneous players for the threshold - I'll leave that to someone else who has a better feel for those numbers.

So clearly there was debate not only about whether to download popular and unpopular games, but also about what constituted *popular*. What level of simultaneous downloads should be chosen? How was Steam calculating and recording this number? How often did it change? The subtext to this disagreement was the difficulty that Melissa had in tracing her purchases in the tumult of transactions generated by the automated downloading. The need for traceability of the financial transactions was in direct contention with the desired level of legibility of Steam games in the RDS. This close examination of Steam purchasing is in fact only a fractal of the complex problems that the NSRL encountered as it collected games from other platforms such as Epic, Origin and Blizzard. While these platforms offer smooth web surfaces for acquiring software, and APIs for the acquisition of data, the logics of financial transactions and the proprietary clients needed for downloading content warped these surfaces, and required the NSRL to develop bespoke software and practices for managing them.

6.2.5. Hashing and Fixity

In the previous two segments I described how collecting activities around Android apps and Steam games worked as appraisal practices for the web. The negotiations needed to perform these software "downloads" were the result of practitioners' necessarily limited knowledge coming into direct contact with the politics of platforms. In this section I examine how the NSRL's focus on *fixity* operates as a representational practice which necessarily shapes its collection. Rather than being fixed and stable, records in the NSRL are subject to constant re-description and re-interpretation which change the nature of what constitutes this archive. Attention to database schemas and data processing activities within the NSRL illustrate how appraisal is not something that happens once, but is part of a "fluid and evolving socially constructed practice" (Yakel, 2003).

A few employees have worked at NIST long enough to have seen the full life cycle of the project and Mike is one of them. He came to the project in 2000 as a distributed systems engineer from another group at NIST in order to help a scientist on the Computer Forensics Tool Testing (CFTT) project with a prototype application that had been built to "hash" software:

> Booz Allen Hamilton had a contract to code up and to build the hashing software that was the very first implementation of the NSRL, and that resided on a squatty little Compaq server that held the database, that had a five-and-a-quarter inch floppy and a three-and-a-half-inch floppy and a CD-drive. In order to hash something, a disk was placed in the machine, the software was installed *on* the server, and the server kept a running track of what was installed on itself and the hashes from the new files got put into the database on that server.
>
> I got there and it was like, you put a disk in and you let it do its thing for about an hour, walked off because it could only do one disk at a time, came back in an hour, see if you could flip the disk, walk off and do something else. So I came, and I was like, there's got to be a way to get the actual work off the server and be distributed, because just having one drive with one floppy every hour is just… You can do better.
>
> So, I went out looking, and was able to find like, I don't know, a dozen or 15 old machines from the excess list that were roughly the same model. I managed to write some code that let it query that database. And it would check and see if that work had been done and if not, it would hash everything that was on the media and store a local copy of anything that it could recurse down into. I never wanted to call it a cluster, because as much as I wanted to do a Beowulf cluster or whatever,

> I never got into really tightly coupling it, so it was more of a constellation.

Immediately at its inception the NSRL was defined by the database which stored the results of the *hashing*. Hashing is an algorithmic process of fingerprinting the digital objects, or files, that constitute the acquired software. Mike helped make this process more efficient by distributing the work of reading the software off of the physical media, *unpacking* or as he says "recursing down into" the data to extract files, and then storing the original media on a shelf in the *NSRL Library* using the Evidence Tracking Identifier (ETID) for shelf-list order. The process of refining the practices of unpacking and hashing have continued to evolve until the present day, and required the creation of *The Corpus*, which is the set of all files that have ever been discovered during processing.

Figure 13: Staff in the NSRL Library, 2014

Almost every NSRL meeting that I attended began with Mike announcing "The

Numbers". These were statistics about how much hashing had been performed in since the last meeting. These statistics were shared in Slack often with accompanying SQL for generating them when people asked questions about recent activity within the NSRL. The Numbers were announced internally when a new version of the RDS was released. For example here is an internal email that Evan, the NSRL systems administrator, sent after he built and released the v2.66 of the RDS:

```
RDS 2.66 September 2019 Hash Counts

Modern:            183,887,293
Modern (minimal):   56,058,194
Modern (unique):    34,149,141
Legacy:             26,202,342
Android:             7,150,516
iOS:                14,390,472

Since June, we have added the following amount of hashes
to the modern sets:

Modern:            +13,278,724
Modern (minimal):   +2,452,001
Modern (unique):    +1,186,763

Great job this quarter everyone!
```

Figure 14: Hash Counts Email

Evan reported these numbers directly out of *Juggernaut*, which is the 4th generation of the NSRL database. The representation of the NSRL Reference Data Set (RDS) itself is produced using a set of queries that run against Juggernaut. The new hashes that are reported in this email are a reflection of new software being added to the NSRL, as activities such as Steam and Google Play Store continue. So for example the addition of 1,186,763 hashes for *Modern (unique)* indicates that 1,186,763 new hashes that have never been seen before were found in the last 10 years of software.

But, importantly, these numbers can also change based on the NSRL's own abilities

to *unpack* software change, and as its own ideas about types of fixity algorithms to use evolve. Mike recalled his initial rewrite of the code that Booze Allen Hamilton had provided:

> We freshly wrote the hashing code, because there was no way to really augment what the contractors had built for us. So it would open a ZIP file, it would open a CAB file, but it didn't know about Java JAR files, it didn't know about all kinds of things, TAR files, anything Unix really.

ZIP, CAB, JAR and TAR are all *container* file formats: they are files which themselves contain other files. Here is how Blake, another NSRL software developer, described the unpacker:

> It says, okay I can treat this box as a box. I don't treat every file I find as a box. So I'm just gonna go through this re-entrant cycle of saying, if I treat this as a box, can I try to tear it open and find things in it? The idea that during the installation process for like 99% of applications that we see, you're just unzipping something and putting it on the disk drive, right?

When software was installed the files in these containers were "unpacked" and written to the computer's storage where the software expected to find them.

It was essential for these container files to be unpacked prior to *hashing* so that the picture of the software be the as complete as possible for identification to later. Today the NSRL unpacking process looks for 28 different container formats. As the NSRL discovers new container formats and learns how to read them, they will rerun the unpacker on the *Corpus* (all the previously unpacked files) to see if new files can be discovered in the files they already have. In addition some software is identified for *Disk Printing*, where the software installer is run in a virtual environment which is later imaged in order to capture files that have been actively pulled from the web

during installation.

All this effort to unpack software containers is in the service of the most important process of all–*hashing*. In his multi-volume encyclopedia of algorithms *The Art of Computer Programming* Donald Knuth describes hashing as:

> The verb "to hash" means to chop something up or to make a mess out of it; the idea in hashing is to scramble some aspect of the key and to use this partial information as the basis for searching. We compute a *hash address* h(K) and begin searching there…A good hash function should satisfy two requirements: a) It's computation should be very fast and b) It should minimize collisions. (Knuth, 1998, pp. 514, 519)

The details and meticulous references in Knuth's 46 page chapter outlining the history of hashing make clear that the theory and practice of hashing is a subfield of its own, that is at the foundation of many concepts in cryptography and databases. Knuth traces the first mention of the concept of hashing back to the early days of computing in January of 1953 at IBM, when Hans Peter Luhn wrote in a memo about the use of "buckets" in information retrieval (Knuth, 1998, p. 547; Stevens, 2018). While many technical improvements have been made since then, particularly for minimizing the number of "collisions" when hashing, the basic concept remains the same. The result of applying a hashing algorithm to content is an index value that can be used to quickly lookup the content, or to verify its content (a checksum).

Hashing was an important concept for the NSRL from the very beginning because the purpose of the NSRL was to compute digital signatures for all the files that comprise software so that these files could be identified later by computer forensics tools. More about the significance of hashing and this identification process will be discussed in the next section. But because the art and science of hashing algorithms was a constantly changing field the NSRL has needed to modify its approach to

hashing over the last two decades.

Figure 15: NSRL Lab Door

In a NIST publication from the same year that the NSRL was officially created Tim Boland and Gary Fisher described the rationale for the initial selection of hashing algorithms for use in the NSRL (Boland & Fisher, 2000). They describe how the computational complexity of calculating hashes using the CRC32, MD4, MD5 and SHA-1 algorithms increases (respectively) as does the robustness of error detection. An error, or collision, occurs when two bitstreams hash to the same value. Errors are usually undesirable in most use cases, and the NSRL is no exception since a collision would result in the misidentification of a software file. When describing the MD5 algorithm Boland and Fisher state:

> The MD5 message-digest algorithm takes as input a message of arbi-

> trary length and produces as output a 128-bit "fingerprint" or "message digest" of the input. It is computationally infeasible to produce two messages having the same message digest, or to produce any message having a given pre-specified target message digest.

The data dictionary created for the NSRL in 2000 lists four hash values as being used, which matches this analysis of checksums by Boland and Fisher (Fisher, 2000). The v1 series of RDS started in 2000 included CRC32, MD4, MD5 and SHA-1 hash values. However in early 2003 a new major version of the RDS v2 was introduced which dropped MD4 support, and coincided with publication on the web (NIST, 2003). The use of CRC32, MD4 and MD5 continues to be used in RDS releases today.

> Although attacks on the various algorithms used with the NSRL are significant for certain cryptographic applications (e.g., digital signatures), in reality they have little impact on how the NSRL is used within the forensics community. If there is a point in time where a given algorithm is deemed unacceptable within the NSRL context, multiple hash algorithms are already included within the NSRL, and NIST has processes in place to easily add additional algorithms as they are needed.

In two versions of the *NSRL and Recent Cryptographic News* page on the NSRL website, which are now only available in the Internet Archive, the problem of collisions with MD5 and SHA-1 was acknowledged (NIST, 2004 , 2017). The pages go on to clarify that while these collisions are deeply problematic for crytpographic applications, they do not present a problem for the NSRL because both algorithms continue to be "pre-image resistant, i.e., it is computationally infeasible for a message to be constructed that matches a given hash." In addition, the NSRL makes multiple hash values available in the RDS, and vulnerabilities in one have not been

demonstrated to affect the others. However, even as early as 2004 plans to move away from SHA-1 were announced, and as recently as 2019 the Subversion revision control system used to record changes to the NSRL code records the database schema being modified to add SHA-256 to the Juggernaut database schema so that it can be released as part of the new RDS v3 series in the near future. The SHA-256 is used to create a unique name in *The Corpus* for every software file that has been discovered.

```
------------------------------------------------------------------------
r3191 | mike | 2019-02-27 11:15:40 -0500 (Wed, 27 Feb 2019) | 1 line
Changed paths:
   M /NSRL/code/trunk/Corpus/unpackcorp.pl

unpacker script bug fix to add SHA256 to HASH table
------------------------------------------------------------------------
```

These details highlight how concerns over hashing algorithms have worked to shape the representations of the RDS, the Juggernaut database, and the NSRL's collecting activities. They also demonstrate why *The Corpus* was needed as a representation of all the known files ever encountered so that they could be easily used as input for new fixity algorithms. The computational complexity of calculating and reporting fixity information from the database impacted its representation.

In 2019 the Juggernaut database had approximately 2 billion rows in its *HASH* table. As the database has grown in size over the years it has needed to be refactored. Refactoring is the process of redesigning software and data structures to suit changing conditions of use. One of the most significant refactorings was concluded immediately before my field study with the NSRL began. The NSRL asked Brian, a database architect who worked on the National Vulnerabilities Database, to refactor

Juggernaut to make it more performant. Brian wasn't told explicitly what the problems were other than queries were taking several days to run. The query in question was the process for generating the RDS snapshot. In addition there were user facing web applications, such as *Collector* which are used to manage the NSRL's activities and had started to slow down and become responsive as they were in contention with the many *unpacker*, *corper* and *hasher* processes that could be running at any given time. Brian didn't know the nature of the exact performance problems, but was able to normalize the database schema to improve performance:

> We never got into that detail…I just applied my basic techniques…what I've done to databases in the past to clean them up and got proper keys built. And I kind of hoped that would solve the problem. Because in the past, that's I mean, 98% of whatever is done when you fix a model and make it right, performance takes care of itself. And there's some reports, you have to add keys and I think we added indexes in that case. Whenever I build a foreign key I will always build an index.

In the process of applying his usual techniques of normalizing database schemas Brian needed to create indexes, which resulted in speeding up the database queries that various processes (RDS generation, web applications, hashing processes, etc) needed to execute. The new performance characteristics significantly affected the volume of software that the NSRL was able to ingest without negatively impacting its operations. It also altered the representation of the acquired software. When I spoke with him, Brian showed me this before and after visualization of the database schema to illustrate the dramatic changes that he made. He specifically pointed out the disconnected layout before, and the connected layout after, as foreign keys and indexes were added to tables.

During my interview with Brian I learned that the reason he initially got in con-

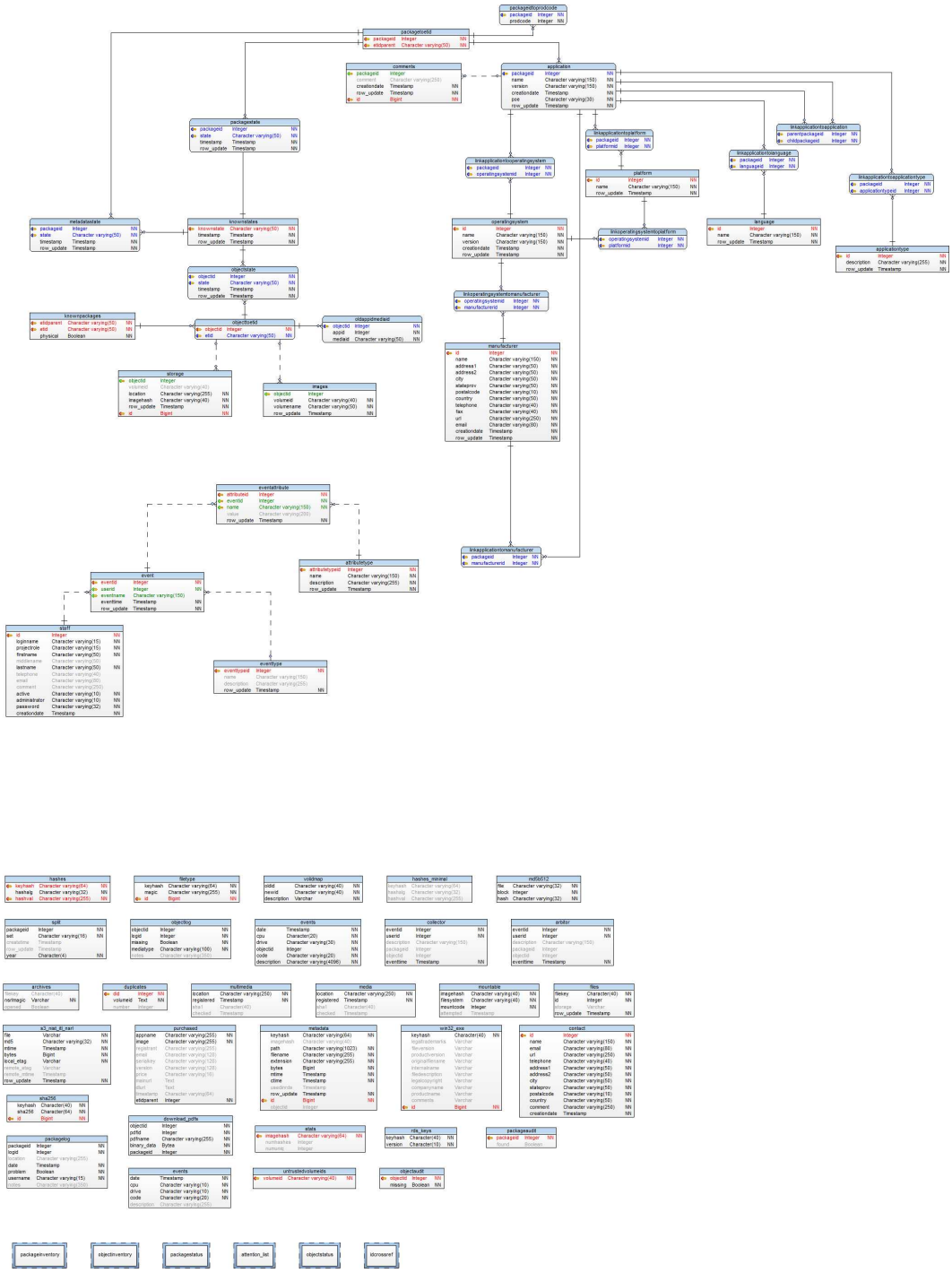

Figure 16: Juggernaut Before Refactoring

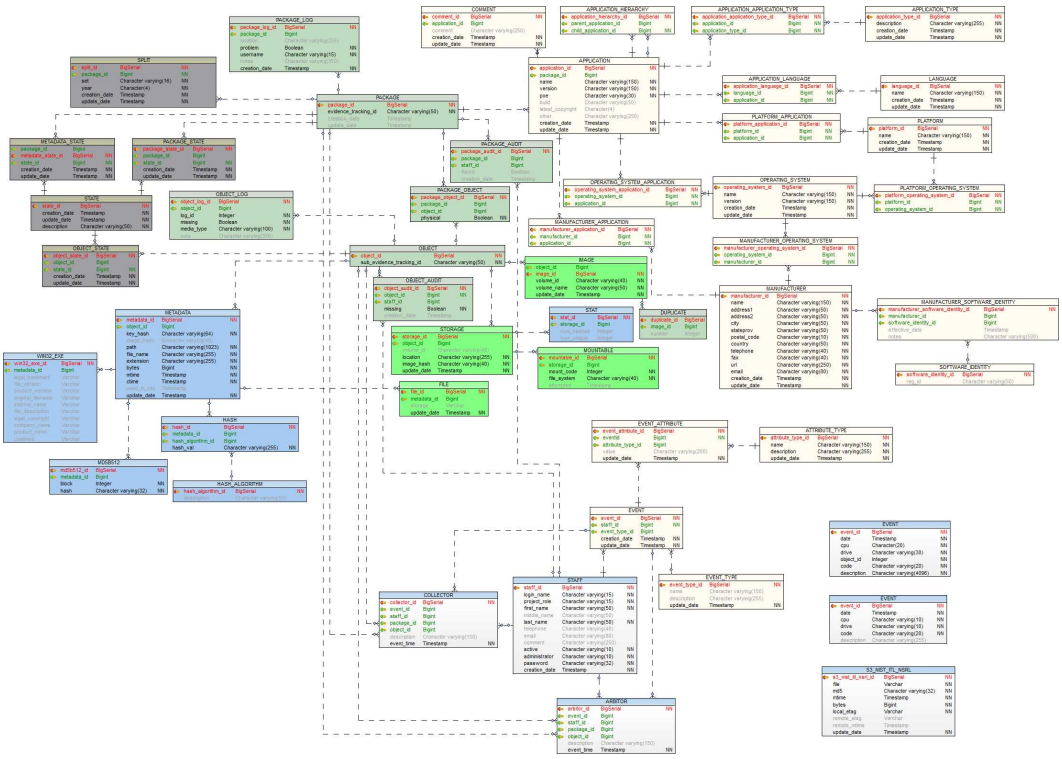

Figure 17: Juggernaut After Refactoring

tact with the NSRL was because he was researching whether the NSRL database (Juggernaut) could be used as a data source to bootstrap the use of Software Identification Tags (SWID) in the National Vulnerabilities Identification Database (NVD). NVD is a public database which provides data for automated management of software vulnerabilities. SWID is a standard for identifying specific software products, which is required for recording which pieces of software are effected by a vulnerability. Brian indicated that the research concluded that the NSRL database was not a good fit for SWID. He did not want to comment on the specifics of why it was not a good fit, but did suggest that I speak to another member of the team about the reason why it was rejected. Unfortunately that person did not respond to my interview request. However, In the process of analyzing the RDS data I did happen upon some aspects of the RDS which I thought could explain at least part of the rationale for not coupling the SWID registry to the NSRL. In analyzing several of the meta-

data made available in several RDS releases I noticed a lack of authority control for manufacturers. For example in RDS v2.69 there were four distinct records for Apple.

Manufacturer	NSRL ID	Software Versions
Apple Inc.	2175	799
Apple Computer Inc.	67	375
Apple, Inc.	82237	3
Apple	65803	1

While two of these are likely the result of an official name change in 2007 when Apple dropped "Computer" from their name, two others appear to be data entry errors. You can see similar normalization issues when looking at other manufactures such as Electronic Arts and Google.

Manufacturer	NSRL ID	Software Versions
Electronic Arts, Inc.	1237	230
Electronic Arts Ltd	310	97
Electronic Arts	40114	7
Google	1618	909
Google LLC	82222	51
Google Inc.	4469	5

These variations indicate that a lack of normalization may be a general characteristic of the NSRL database on the whole, at least when it comes to manufacturers–but perhaps also to other entities as well. It is important to stress here that these are

not *errors* in the usual sense, because the NSRL's database and practices were designed for collecting, unpacking and fingerprinting software, not for allowing software manufacturers to register their software releases in a canonical registry like SWID. The choices of representation made in the design of the NSRL made some use cases possible, while precluding others. It's conceivable that automated collecting from the web, such as the Android and Steam collecting examples above, could swing the pendulum back in the other direction, by regularizing the ways that software is collected and recorded, which could provide at least the appearance of a cleaner line of provenance.

6.2.6. The Customer

In the final four segments of my findings I examine different examples of *use* of the NSRL. For example the story of the NVD and the NSRL is a story of failed use, or perhaps even *disuse*. My findings thus far have focused on the internal workings of the NSRL, but will now turn to examining the broader significance of the NSRL by tracing its activities out into the field. One way of doing this that I found useful was to follow the RDS, and its millions of hashes as what Bruno Latour calls *immutable mobiles* or *inscriptions*:

> A general term that refers to all the types of transformations through which an entity becomes materialized into a sign, an archive, a document, a piece of paper, a trace. Usually but not always inscriptions are two-dimensional, superimposable, and combinable. They are always mobile, that is they allow for new translations and articulations while keeping some types of relations intact. (Latour, 1999, pp. 306–307).

It wasn't long after I started attending NSRL staff meetings that I heard "The Customer" being mentioned. For example, in one meeting when team members were

discussing how many new hashes would be made available in the next RDS release, Lara, one of the NSRL's managers reminded the team that "two million new hashes isn't what matters; what matters is that the release is driven by what the customer needs". James and Mike responded saying that the new hashes for video games would be a significant addition to the release. Melissa, who is responsible for purchasing software for the NSRL, followed up by saying how she "always tries to think like a customer, and how I've done a good job with the customers". This made me wonder who the *The Customer* was, and what it meant to "think like a customer"?

The generic term "customer" wasn't used by team members to refer to all the users of the NSRL, but only to a certain, very important, set of users of the RDS. The NSRL's website clearly documents that these users are members of the law enforcement:

> The National Software Reference Library (NSRL) is designed to collect software from various sources and incorporate file profiles computed from this software into a Reference Data Set (RDS) of information. The RDS can be used by law enforcement, government, and industry organizations to review files on a computer by matching file profiles in the RDS. This will help alleviate much of the effort involved in determining which files are important as evidence on computers or file systems that have been seized as part of criminal investigations.

It is important to note that the NSRL was conceived in 1999-2000 at the exact same time, and by the same set of actors, who created the Computer Forensics Testing Tool Testing (CFTT) project. In an article published in NIST's *ITL Bulletin* shortly after they were created, Gary Fisher, the first manager of NIST's digital forensics projects described the NSRL and CFTT projects together as part of an orchestrated effort. Both projects were conceived, articulated and funded by NIST in collaboration with the National Institute for Justice (NIJ), the Federal Bureau of Investigations

(FBI), the Department of Defense Computer Forensics Laboratory (DCFL), and the Department of Justice Technical Support Working Group (TSWG). The CFTT establishes testing practices for digital forensics software: disk imaging, write blockers, disk wiping, deleted file recovery etc. These tests are then applied to commercial and open source digital forensics products by CFTT project members, and the results were published on the NIJ website.

The RDS and CFTT were articulated together in order to help courts decide whether expert testimony about digital materials should be admissible as evidence. This is known as the *Daubert Standard* established by the Supreme Court decision in Daubert v. Merrell Dow Pharmaceuticals in 1993 which indicated the four criteria that a trial judge may use to assess the admissibility of expert witnesses' scientific testimony during federal legal proceedings:

1. Whether the theory or technique had been tested.
2. Whether the theory or technique has been subjected to peer review and publication.
3. Whether there is a known or potential rate of error and whether standards exist to control the techniques operation.
4. Whether the technique has general acceptance within a scientific community.

The CFTT tests and reports help address these concerns as applied to digital forensics tools. It is no coincidence that the NSRL RDS is used by many of these forensic tools to filter out known files, so that investigators can focus on in on the unique ones. *The Customer* are the operators of these tools.

The CFTT reports started publication on the NIJ website in 2002. It is interesting to note that the language which describes but does not fully delimit the set of actors that were involved in its funding: "CFTT is supported by other organizations, *including* the Federal Bureau of Investigations …". (DHS, 2020, p. 5, emphasis

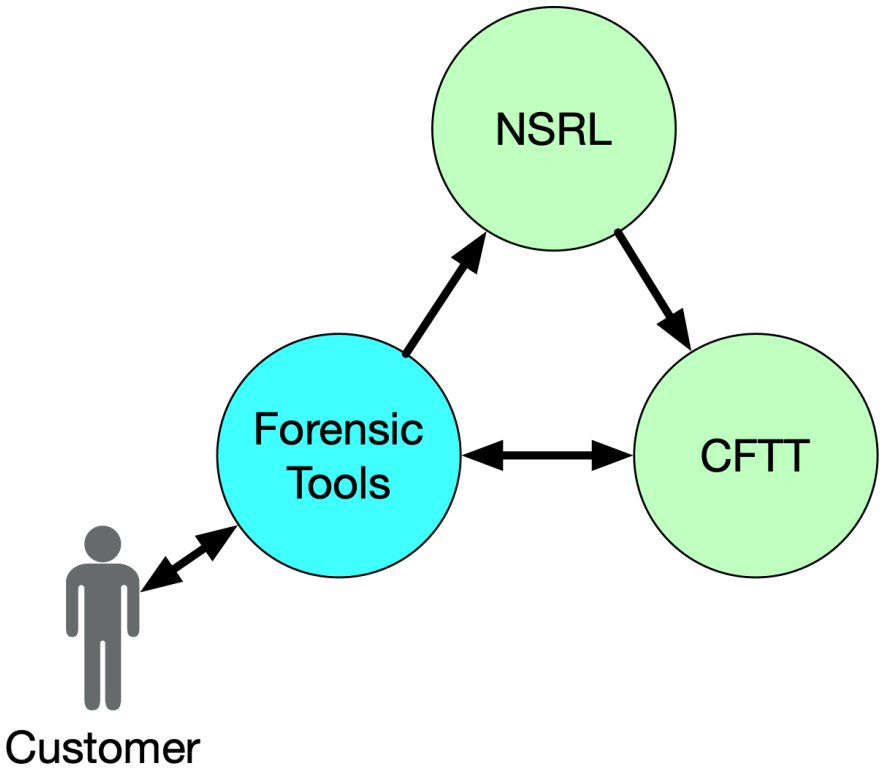

Figure 18: The Customer

mine). It wasn't uncommon to hear staff talk in passing about "certain three letter organizations" implying, but naming the Central Intelligence Agency and the National Security Agency, who would clearly also have an interest in digital forensics. Even as early as 2002 NIST's Information Technology Lab's annual report talks about the significance of NSRL and CFTT for combatting terrorism:

> Since the tragedy of September 11, 2001, ITL's work in computer forensics has taken on added significance and impetus. Sound computer forensics practices are a key to finding and delivering court-permissible evidence when computers are used in the commission of a crime. (NIST, 2002a, p. 12)

Team members generally agreed that "thinking like a customer" meant thinking like law enforcement using a digital forensics tool to "DeNIST" a set of computer files.

Doug White and Mark Rose (seated) help the law enforcement community apprehend criminals and terrorists by developing reference data and tests for computer forensics.

Figure 19: NIST (2002)

DeNIST was a term of art for applying the digital signatures in the RDS as a sieve to the digital signatures of files on some computer storage, in order to locate the unique files, or files that weren't part of known software packages. For customers performing DeNISTing the more generic files they are able to eliminate from the investigation the better. Recall Figure 1, which continues to be used to demonstrate the value of the NSRL in slide presentations to this day. All the staff agreed, the best way to achieve the highest percentage of matching was to collect "popular" games for popular operating systems, or as Lara told me:

> So many people are using it [the RDS] just to get rid of known software. That's our number one use case, to support the efficiency of these investigations, by getting rid of known content so you don't have to search it. For that, for everybody, the metric is popularity. Have big popular things.

As I spent time with NSRL staff I learned about the *Steering Committee*. This com-

mittee has met every 3 months for the past 20 years, to help guide and provide advice to the CFTT and NSRL projects. Membership on the committee is limited to federal law enforcement agencies. The first version of the NSRL website available in the Internet Archive collected on March 2, 2001 has a page for the Steering Committee listed in its menu. Unfortunately the page itself was not archived at this time. The next version that is available from April 1, 2001 does not include the Steering Committee page, and the name only appears briefly with no details on the NSRL website after that. When I spoke with her Lara indicated that they needed to keep the committee membership limited to federal government to reduce the amount of time and effort it would take to organize the meetings:

> The Federal Advisory Committee Act means you have to have open public meetings. You have to announce stuff 90 days in advance. There's a lot of overhead. So if the government's getting advice in from the public, there's a more formal process to make sure it's fair and open. It's not a bad thing.

So rather than solicit requests for software directly from the users of digital forensics tools, or their manufacturers, these requests for software came from Steering Committee members, who worked in the field of law enforcement. During my time with the NSRL I only spoke to two members of the NSRL team who had attended these meetings. In many ways *The Customer* became a way to talk about *The Committee* in conversation, since *The Committee* was a proxy for the actual customers, who they couldn't get formal advice from directly. During my time with the NSRL I only remember hearing about the committee making one request: improved Windows 10 coverage, which got discussed in Slack:

James @Mike Windows 10 complaints?

Mike	Yeah, Lara got some DHS feedback about "NSRL is so behind publishing Windows 10 data that some organizations are making their own W10 NSRL's to make up for it"
Melissa	Oh woah I don't think I knew this! I just knew we should focus on Windows stuff. When was this?
Mike	Like 6-8 months ago. we're on it now, or at least more than we were.
Mike	probably time to do a field test, hash someone's W10 laptop and see what RDS covers

These requests came in while the NSRL was hard at work keeping their collecting from platforms like Steam and the Google Play Store going while they shuffled their processing around after failure of one of their primary storage systems. A few weeks later Mike followed up with the results of the test:

Mike	@Melissa bad news on the Windows10 front so far (about 25% through checking) - we're only identifying about 12% of the files. There were 139,500 distinct SHA256 hashes found on that PC, I've checked the first 25% of them, 34,807 of them, 30,400 (87%) are unknown by the NSRL database. (about 6,000 are user's files) Here are the directories with the most unidentified files …

This specific example demonstrates the direct connection between use of the RDS and the appraisal activities of the NSRL. Microsoft Windows is arguably one of the most popular pieces of software in the history of computing. And yet the question ifo whether or not Microsoft Windows has been ingested fully into the NSRL is constantly open to negotiation and revision.

6.2.7. CAID

While *The Customer* and the *Steering Committee* are certainly important for understanding the dynamics of appraisal in the NSRL, their activities remained abstract and remote. I lacked insight into the decisions that had been made over the years. Fortunately a singular event occurred during my time at the NSRL which greatly aided me in seeing the network of actors that participated in the NSRL's appraisal decisions. This segment highlights how NIST's own storytelling about its work helped them articulate and instantiate the types of values that motivated the NSRL's work.

One of the most significant events that I witnessed during my time with the NSRL was an all hands staff meeting that was called in December of 2018. The meeting wasn't announced beforehand as an all hands meeting, but on entering the room I immediately knew that this meeting was different because I could see that the contractors who worked on downloading software were present. Additionally I saw some unfamiliar faces of NIST staff, who I later learned were from the Computer Security Division (NSRL was a project of the Software and Systems Division). In addition Lara, one of the head managers of the NSRL, and her immediate supervisor Frank were there. Lara occasionally came to the NSRL staff meetings, but I had never seen Frank in one of the meetings.

I was sitting between Lara and Mike who filled some time by chatting idly about *The Numbers*: 80 million hashes had been added, 2 million were distinct. Lara also announced that NIST was starting a new "black box" research program to study digital forensics practices. James suggested that his recent research on detecting file systems could be of some interest. It was clear that Lara and Mike were waiting for others to arrive before starting the meeting. After a few minutes the Director of NIST's Information Technology Lab entered the room and sat down at the head

of the table. Once everyone had exchanged greetings things got quiet and Mike welcomed everyone.

> So I'd like to tell you why I've called you all to this meeting. As you know we don't often get feedback about why we are doing the work we do here in the NSRL. Such as when we were able to make word processing software available to the FDA; when we helped Nobel Prize Winner Bill Phillips recover a very important file that was part of his research; the work we did with Stanford on the Cabrinety Archive … and the work we did with the Child Abuse Image Database. It's this last one that I want to talk about today. I just want to congratulate everyone on the work you did to get Blizzard, Steam and Origin into the RDS. Just a few weeks ago Lara and I were in the UK for a conference and Mark and Francis at the Norfolk Constabulary, and the Home Office of the UK have commended the work of the NSRL. The Norfolk Constabulary has made a gift to the NSRL, an honorary Norfolk Constabulary Helmet–I was told that this was the first time a helmet has been awarded.

Mike took the helmet out of an empty box of veggie-chips, removed it from a black cloth bag and passed it around the room as he joked about it not containing Guinness. As the helmet was passed around the room I remember feeling like this was an opportunity for each member to hold a physical artifact that signified the importance of the NSRL. People laughed as the director joked that the helmet was the best piece of swag anyone had ever gotten at a conference. He then went on to say how significant the work the NSRL was, and that it was impossible to measure the impact of helping save children and this database that had been created.

The helmet seemed exotic and strange, an artifact from another time and place. But

it also seemed familiar, at least to me. I remember smiling as the helmet was passed to me, and as I held it. It felt heavier than I imagined it would be. I examined the insignia, and was reminded of my own childhood in England when I had a toy version of such a helmet. When I wrote up my fieldnotes for the event later I paused to consider how many people in the room might have similar memories. What myth making was this small ceremony doing? It reminded me of the familiar slide from Figure 1 with Sherlock Holmes performing an investigation. Everyone gathered after the helmet had been passed around the room and stood up against the pale cinder block wall of the meeting room for a group photograph. I don't remember a copy of the photograph circulating. I was told later that the story was going to be published on the NIST website, but I wasn't able to find it. The short write up prepared for the website was distributed as a Word document in the NSRL Slack. A year later I was in the NSRL JNet lab room and noticed the helmet up on a shelf near the ceiling above one of the contractors' desk, who was responsible for manually downloading software which could not be downloaded automatically.

When I asked Mike later during an interview about how the NSRL had started collecting games he had told me the story about how the CAID collaboration came about.

> So I had been in contact with some people in the UK who run the Child Abuse Image Database, CAID, and they approached us and said, "How much of this gaming software do you have? We want to try and collect all of the benign images in multimedia." So then they go after GIFs and JPEGs and M4As and all of the multimedia type things to try and winnow those off of a system to filter those out so that they can focus on any of the child abuse images that might be left on the system. And they said, "Oh, we can contact Blizzard, and can you guys contact Steam

Recognition of the contribution of the NSRL to helping law enforcement catch child predators

In November, Doug White, the PI for the National Software Reference Library, went to the UK to meet with the Child Abuse Image Database (CAID) team. CAID is run out of the UK Home Office with involvement from constabularies around the country. The CAID project helps identify and safeguard victims, helps improve the investigation of child sexual exploitation and abuse, and supports international efforts to remove images from the internet.

The CAID team presented [redacted] with an official ceremonial Norfolk Constabulary helmet in recognition of how helpful the NSRL is to their efforts. (The chief implementation officer for CAID is based in the Norfolk Constabulary.) In particular, they cited the inclusion of gaming software to the NSRL as helping with their efforts. As far as the CAID team knows, this is the first time a ceremonial helmet has been presented to someone outside the UK.

The NSRL has been supporting CAID since 2014 and other efforts to stop child exploitation since we started the project in 1999. Just this year the Southern Virginia Internet Crimes Against Children (ICAC) Task Force started using CAID in their efforts.

Figure 20: CAID Recognition

> because they're in the US?" And I said, "If you'd like to collaborate with us on this and let us do the collection, we've got the infrastructure to hash all of this stuff and we publish it publicly, so if we combine our forces to lobby these game manufacturers to send it to us as just one point". Because they were looking at, if we can get this game and somebody else can get that game… and it's just like, it's easier for a company to know, "All we have to do is funnel it to the NSRL," rather than worry about the contacts for various and sundry places or other law enforcement organizations. And we publish it publicly, so it would absolutely come out. So that's how that whole gaming concept got started, it was the UK.

Hearing Mike describe this relationship and how it evolved made me realize how significant this network of relations between users of the RDS (CAID), the RDS and NSRL (NIST), and software manufacturers (Steam and Blizzard) have been to the

NSRL's appraisal practices. It made me consider why it was that this distant story of use was told, rather than ones closer to home that we could imagine involving the "three letter organizations" such as the FBI, ICE, CIA or NSA.

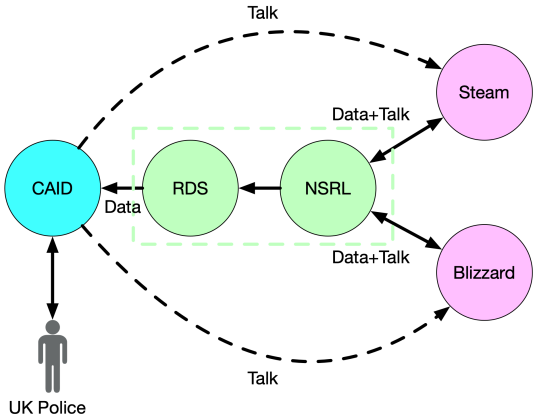

Figure 21: CAID and Games Archiving

6.2.8. ByLock

It seems obvious that these stories of use in the NSRL cohered, or hung together, in some fashion, because they all involve a network of actors that includes the NSRL in some central capacity. But even though they cohered, it became apparent as I reread and coded my fieldnotes that these stories were not, taken as a whole, coherent. The uses for the NSRL that I was witness to demonstrated how the NSRL is constituted as a heterogeneous set of actors that are better conceived of as an *assemblage* with multiple shifting functionalities (Deleuze & Guattari, 1987), rather than as a single fixed entity. A project of the size, scope and duration of the NSRL accumulates many stories of use over its lifetime. In this segment I explore a story of use that was hardly told at all, but which highlights a divergent, or perhaps even unwanted use (disuse), and turned out to be a use at all–at least in the conventional sense.

A staff meeting a few months earlier was the scene for a much less auspicious announcement of NSRL use. It took place in the space of only about three minutes,

but the exchange seemed significant enough for me to jot it down in my notebook, and write about it more fully in my fieldnotes later that evening. During my analysis of field notes and interviews this brief moment presented itself as an alternate story of NSRL use.

We were going around the room round-robin style as we usually did. Evan, who did much of the systems administration for the NSRL, had just announced that Rack A had become unresponsive, again. This was a piece of hardware that provided much of the storage for the NSRL, and which had been causing problems since I began attending meetings three months earlier. After Evan and Mike finished talking about how they could bring back some of the storage using a FreeNAS server the updates went to James. James is a software developer and computer scientist, who had done his PhD dissertation while working as a Guest Researcher with the NSRL.

James announced that he had attended the Open Source Digital Forensics Conference the week before, where he had heard from people who were using the XML standard that he had helped develop. He continued somewhat sardonically by saying, "Oh, and apparently Turkey is using the RDS?" Mike raised his eyebrows at this, and the others looked surprised as James continued, "Yeah, someone gave a lightning talk, and described how he worked for the Turkish government, which had used the RDS to inspect 1.6 million devices during the investigation into the 2016 coup attempt, which led to 75,000 people being put in jail?" Lara, one of the NSRL managers commented "Great…NSRL data being used to prop up military dictatorships." To which Mike responded "Yay us?", and then the meeting moved on after a pause.

James, Lara and Mike's sarcasm here underscored their critical perspective on this use of the NSRL. A few years earlier on July, 15 2016 a faction within the Turkish Armed Forces had attempted a coup to remove President Recep Tayyip Erdoğan and

his government. The coup against Erdoğan failed, and ignited a series of purges that have led to hundreds of thousands of government employees losing their jobs and others being imprisoned. The Turkish intelligence service alleged that individuals had used a mobile application called ByLock to coordinate the coup, and installation of the app on mobile devices was used as evidence to arrest individuals (Gokce, 2018).

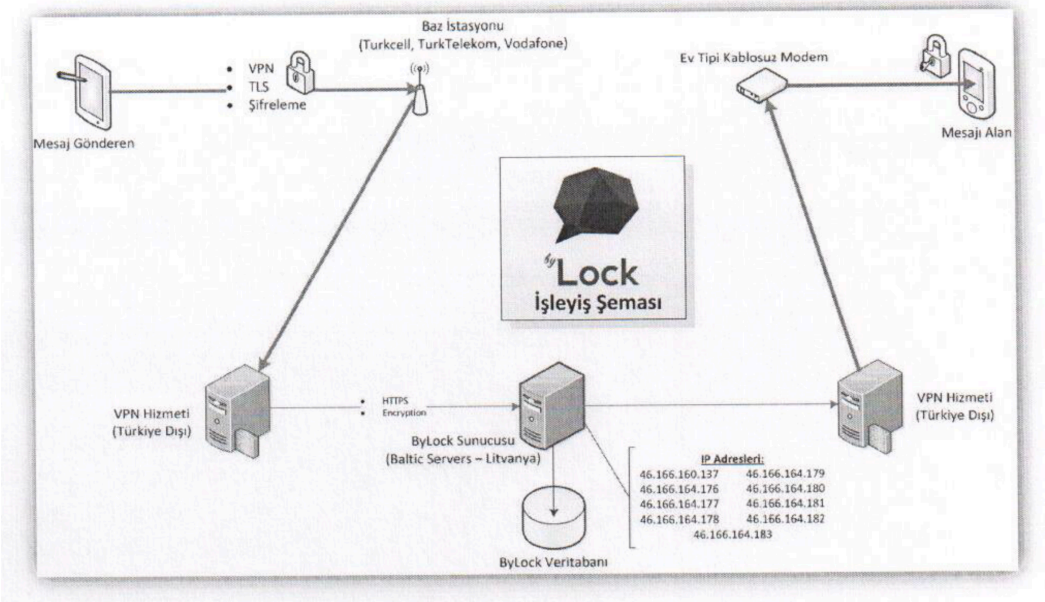

Figure 22: Turkish Intelligence Forensics Diagram from Fox-IT (2017)

The NSRL employees clearly understood the significance of the claim being made in this conference presentation. As a public dataset the RDS can be used by anyone who has an Internet connection, and the technical skill to operate tools that use it. These skills are not difficult to come by. The ByLock story highlights how NSRL employees are well aware that the RDS is a *dual-use* technology. Dual use is usually used to describe technology such as the Global Positioning System (GPS) which can be used for both civilian and military operations. However the NSRL RDS is dual-use in more ways than one. The RDS can be used to "DeNIST" computers, as demonstrated in the CAID use case where the RDS to focus their analysis of child pornography by removing known files. But the RDS can also be used to *identify*

where software has been installed, by looking for fixity matches across a device's file storage. The RDS is dual use in another sense in that it can be used by the US government, and the Turkish intelligence services, and conceivably by criminals or terrorists themselves. NIST refers to the RDS as "neutral" but in actuality it is more accurate to describe it as ambivalent. Whatever the case it's a good reminder of Melvin Kranzberg's insight that "Technology is neither good, nor bad; nor is it neutral." (Kranzberg, 1986).

Ironically, I was not able to locate the ByLock application in any of the product listings included in recent RDS releases. I was only able to check back to RDS v2.58 (September, 2017). I was also unable to find a name for the presenter on the Open Source Digital Forensics Conference website, which isn't unusual with lightning talks which tend to be impromptu. I followed up with James and the conference organizers to try to find more information about the presenter but was unable to get any leads. The whole episode left me feeling queasy, as I wondered if James had heard the presenter wrong, or one of them wanted to portray the Turkish government as a user of the RDS for some reason. I found myself considering how some could even use this as an argument for closing access to the RDS, much like access to the software in the NSRL is only available to people who have been granted access to the NIST campus, and to the NSRL itself.

6.2.9. Cabrinety

In this last story of use, and the final segment of my findings, I want to continue to look through the lens of *use* at the heterogeneous sets of actors that participate in the NSRL's appraisal activities. Unlike the ByLock example, this story of use is openly celebrated by the NSRL, through publicity announcements, interviews, and the long term engagement by multiple NSRL staff. In it we see again how important institutional collaboration is for appraisal. But in this case the immutable mobiles

we will trace are not the millions of hashes, but the millions of files that comprise the software being archived.

Many of the conversations I was part of with NSRL staff during my time there seemed to revolve in some way or another around their active work collecting games. So I wasn't surprised at all when one day after logging into NSRL Slack I saw a conversation thread about the recent release of classic interactive fiction text games that were originally published by the company Infocom (Axon, 2019). An archivist working at the Internet Archive named Jason Scott had received an "anonymous donation" of the source code for these Infocom games, and placed them all on code sharing website GitHub (Scott, 2019).

In the #general Slack channel Mike shared a link to an article written by Samuel Axon in Ars Technica about the release of the software by Scott and went on to discuss its significance for "historians, narrative designers, programmers and game enthusiasts". Mike asked other channel members if the NSRL should collect the source code? James responded quickly saying "Just clone it". To which Melissa pondered "I wonder if these are part of the Cabrinety collection." I had heard the name "Cabrinety" in meetings before so I asked Mike about it later when I saw him. He said that the NSRL had acquired disk images of software from Stanford University, but that it had been four years since they received them, and they still had not yet added the hashes for the software to the RDS. "But they should" he added.

The Cabrinety Collection at Stanford University is an archive of microcomputer software originally assembled by Stephen Cabrinety between 1975 and 1995. Cabrinety began collecting software in high school, and continued to add to the collection throughout his life until his untimely death at the age of 29. In 1982 he had dropped out of Stanford University to found Super Software Inc, which produced educational

software. In 1989 he founded the Computer History Institute for the Preservation of Software, which was the first non-profit organization dedicated to collecting software. When he died there were approximately 50,000 software titles in the collection. Stanford kept the collection stored in an off-site, climate controlled warehouse with limited access, and published a finding aid in 2000 (University, 2000).

In a podcast interview for Crime and Science radio Doug White, the then and current project director of the NSRL, described how a chance meeting in 2009 between Michael Olson of Stanford University Libraries and Simpson Garfinkel, a digital forensics expert then at the Naval Postgraduate School, led him to pay Stanford a visit in May of 2009 to learn more about the collection from Olson and the collection's curator Henry Lowood (Lyle, 2017). Here is his description eight years later of this initial visit and what followed:

> White: I had a laptop, a floppy drive, and some extraneous floppy disks, and I showed Michael and Henry an example of our [NSRL's] capabilities. Michael and Henry showed me a few of the historically significant software and hardware items in the collection at the time. It was a wonderful room to walk into–just taking a step back 25 years into history and seeing some of the old equipment and some of the wonderful old packages they had on site for teaching some of their courses. So after we evaluated the benefits, the risks, and the costs of NIST applying the NSRL processing to the Cabrinety software media Stanford University Libraries applied for a grant from NIST in 2012, they were awarded that grant in 2013, and the rest, as they say, is literally history.

This initial meeting was clearly an opportunity to share knowledge and expertise as well as much as it was a way to explore the sharing of software media. In an interview I conducted with Lowood in 2020 he described the Cabrinety collection

as the "gift that keeps on giving" because of the various projects that it has been at the center of over the years. When I asked him about NIST's interest in the collection he said that the age and state of the software was quite important to NIST.

> Lowood: The Cabrinety collection helped them with a problem they were having in completing their collection. They like to have software that has not been used, not been broken out of the box, which is very difficult to do for older software. So they didn't have very many signatures for older software. Doug picked up on that and contacted us to see if we would be interested in the collaboration and it fell into place from there.

The provenance of the physical media was a key factor for the NSRL, since the NSRL was used to generate the RDS in order to identify known software files, in order to eliminate them from a forensics investigation. Seeing the closed boxes was an indicator that the media had not been written to, tampered with, or used in any way prior to the imaging that NIST was to perform. When asked by the interviewer what it was like for White to start receiving the items from the Cabrinety collection he recalled:

> White: Breathtaking. Literally. The first box that was shipped to NIST was one of the archival boxes taken straight from the warehouse, placed in a protective shipping box, and delivered to us. When it arrived I took it to our secure facilities, checked for damage, and opened the outer box, and opened the archival box, and was face to face with twenty or so shrink wrapped titles from the 1990s. I couldn't believe my eyes. I closed the box, telephoned Henry to confirm that Yes, NIST should break the seals, and go to work, and I'm sitting here, I still get goosebumps telling this story. It's as close as you can come to archaeology,

> and finding a sealed tomb or something. Just amazing. I'm perfectly comfortable admitting that there were certain titles that sent me reeling with nostalgia…The Cabrinety collection always had something to challenge us, sometimes it was an odd physical piece of media, sometimes it was a previously accounted file type, sometimes it was an odd metadata relationship, we didn't always overcome the challenges but it was fun to try.

White's description is tailored for the Crime and Science Radio podcast, which through its 67 episodes over 4 years, explored stories of forensic science for an audience of writers and others who were interested in the presentation of forensics techniques in the arts and entertainment. The podcast was hosted by two crime/mystery writers: DP Lyle (also medical doctor) and Jan Burke. Lyle and Burke have consulted on television programs such as Law and Order and CSI Miami, while also advocating for the advancement of forensics science in venues such as the American Academy of Forensic Sciences.

Matthew Kirschenbaum's concept of the *forensic imagination* is useful here for identifying this elusive inter-disciplinary surface, or site of genre collapse. He writes "forensics is commemorative as well as juridical, and fundamental to the arts as well as the sciences" (Kirschenbaum, 2008, p. 250). Others such as Amelia Acker have used this idea of the *forensic imaginary* to describe the ways that every day use of mobile devices enacts the circulation, display and deletion of records as they move through networked infrastructures (Acker, 2015). An *imaginary*, something that lives in the imagination and not in reality, seems at odds with the goals of forensics which aims to use the tools of science to approach reality. The *forensic imaginary* is able to keep these two definitions from pulling apart by drawing us into the creative and emotive work of inquiry. White's emotional response to this initial unboxing is

emblematic of the key role that *affect* plays in the creation and use of archival work, and indeed, appraisal (Cifor & Gilliland, 2016).

However, I wanted to highlight this quote from White, to examine the way this forensic imaginary, or affective experience, is put to work. White's telling of this story sets the scene of the initial unboxing/accessioning of the Cabrinety archive, but it also sets in motion and valorizes a circulation of practice. To illustrate here is an example of Lowood describing what he saw as the benefits of the Cabrinety collaboration:

> First and foremost we were able to get thousands of titles from the collection off the original floppies, and data tapes into formats, disk images that we could put into the Stanford Digital Repository, into portable preservable objects. That was a big thing for us. That opened up some other projects. That opened up the possibility of access to the collection in different ways. It was something we knew we would have to do anyway, and to get that through the project was really great. It was the jump start for some of the projects that came after that, such as one we are working on now, Emulation as a Service, where having the disk images so we can immediately work off a disk image, instead of going title by title that we want to use, and go through some reformatting process in house title by title, which would be very laborious. It was a fundamental thing we needed to do.

On the NIST side, expertise was gained while dealing with new media types, file formats and descriptive metadata as it became an active participant in discussions about digital preservation and the preservation of software. On the Stanford side there were the clear benefits of getting disk images for their software that lay dormant in a warehouse, while also learning about the application of forensic technolo-

gies such as disk imaging, write blockers, digital object packaging, in an archival setting. This circulation of forensic practice can be primarily thought of as a discursive movement, where knowledge and power are reproduced together (knowledge-power), as the disciplinary contexts of defense/intelligence gathering/law enforcement fuse with archival practice and historical inquiry.

So far in my findings the NSRL has been used to create the RDS for forensics investigations. But the Cabrinety use case seems different because even after four years the files for these historical pieces of software had not been added to the RDS. Of course it makes sense that there has been little pressure to add them since at the time of the reformatting in 2013 the newest software titles from the Cabrinety archive would have been 18 years old. Indeed 80% of the Cabrinety collection itself was published in 1988 or before, which would have made the software 25 years or older at the time it was acquired by NIST. This is not popular software, in fact it's the exact opposite, it is *rare* software. The likelihood of finding matches for this software in the process of forensic investigations is close to zero, unless the subject of investigation happened to be an expert in the history of computing. Maybe there's a good idea for a novel here, but it's not a novel idea for forensics investigations.

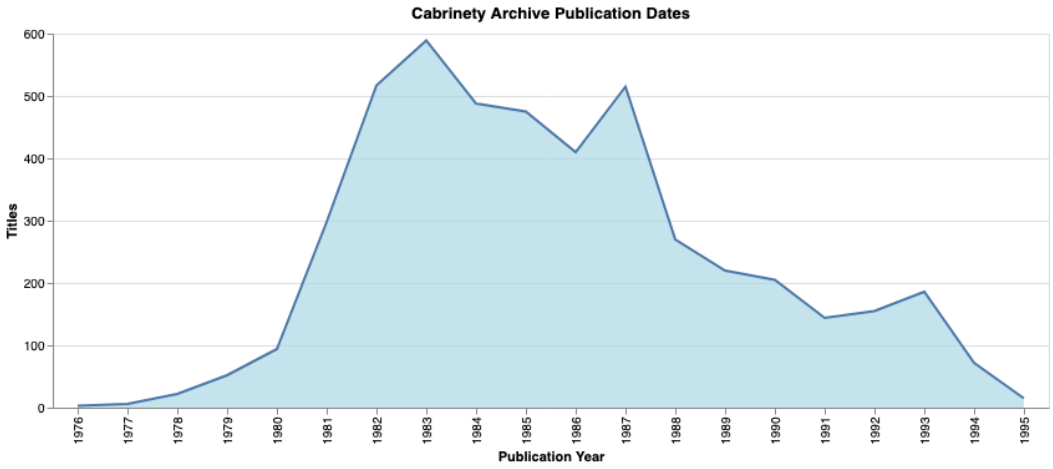

Figure 23: Age of Cabrinety Materials

Taking a step back from this paradox (the collection of rare objects when popular ones are needed) while recognizing the clear affective forces at work while acquiring and describing these historical materials helps to bring the NSRL's appraisal activities into sharper focus. The disciplinary fusion of forensics and archival science achieved in the NIST-Stanford collaboration made it a worthwhile endeavor for NIST, because it was helping to spread a practice and form of *knowledge-power*. The NSRL is part of NIST, whose mission is, after all, the dissemination of standards:

> To promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life.

I learned near the end of my study that not only were practices and technologies circulating as a form of *knowledge-power* but the software artifacts themselves also take part in this circulation. Stanford is a participant in a multi-institutional collaboration hosted at Yale University called the Emulation as a Service Infrastructure project (EaaSI). The ambitious goal of this project is to build a software platform that will allow network of institutions to share software and operating systems to make them usable for historical purposes. After prefacing that he wasn't entirely sure it had happened, Lowood shared with me how he had heard some of the Cabrinety materials circulated:

> My understanding is that some of the objects that we are using in this environment [EaaSI] that were provided by Yale as the PI for this are actually objects that came from the NIST copy of some Cabrinety titles. So we are actually at Stanford through this environment via Yale via NIST circling back to Stanford–from the original project a decade or so ago…It is interesting that we are seeing instances of items from our

> collection through this circuitous route made possible by this project with NIST.

Putting aside doubts about whether this circulation happened or not, it is significant that the same Software Preservation Network that promotes the EaaSI project has also been directly engaged in establishing a legal precedent for allowing this circulation of software to happen under the copyright provisions for Fair Use (Aufderheide et al., 2018). As Lowood described this to me I remembered that during my first few weeks at NIST Lara mentioned this report by Aufderheid to me, while remarking on its significance. An opening for the legal sharing of software with users was something that NSRL management had a close eye on because it could potentially open up the NSRL to a whole set of users that lay outside of its direct mandate to build the RDS for forensic investigations. This example of the Cabrinety archive highlights how fields that are as seemingly unrelated as the military industrial complex and cultural history combine in a particular mode of knowledge-power to create new practices.

6.3. Discussion

At this point it is worth recalling the research question that I began this study with: what are the sociotechnical factors that influence how content is appraised in web archives? During my year observing activities at NIST I discovered that web archiving processes can take different shapes than the ones we are accustomed to seeing in the service architectures provided by the Internet Archive and national libraries. My findings show that the NSRL was actively engaged in collecting web content, specifically software from web accessible software distribution platforms. However the NSRL's practices for collecting software from the web grew out of decades of experience collecting software from physical media. So it is hardly surprising that they chose to continue to cultivate their own methods and tools instead of us-

ing "standard" web archiving tools that implement (*The WARC Format 1.1*, 2017) (e.g. Heritrix and Wayback Machine). These findings also suggest that looking at web archiving practices this way significantly enlarges and deepens the scope of activities we need to be studying as information science scholars.

When specifically considering appraisal practices in web archives my findings elicit three broad areas of sociotechnical interaction. The first is *use* which refers to how the records are put to work shapes what records are created. The second area is *legibility*, or how representational capacities work to shape what is archived. And the third is *governmentality* or how appraisal can be see as an expression of a particular disciplinary mode of rationality. At the risk of over stretching a metaphor I contend that *use* and *legibility* are two sides of the same coin, whose denomination is *governmentality*. To close this segment I will summarize how these findings illustrate how *use*, *legibility* and *governmentality* are interrelated.

6.3.1. Use

There is a red thread of *Use* that can be traced through the NSRL findings above. Consider the collection of software that we saw described in Darius' efforts to download Android apps from the Google Play Store. Or how Melissa, Vlad and James worked together to collect video games from the Steam Platform. At first glance these activities seem divergent: some of the software being collected was commonly used messaging apps, some were popular video games, some platforms required writing software to crawl specific web pages and interact with APIs, others required purchasing gift cards at CVS, and negotiating informal license agreements with the Steam and Blizzard platforms.

All these activities were ostensibly governed by a driving principle to collect *popular* software, so that the RDS would contain *more* digital signatures that would make

it *more useful* to forensic investigators. NSRL staff expressed an explicit value in collecting popular items, even as the metrics for popularity were contested, and escaped measurement due to the opacity of software distribution platforms. As long as there was general consensus that what was being collected was popular the work could proceed. Even software that was specifically requested by *The Customer* was done so for forensics investigations, where the appraisal value was clear. This line of reasoning about popular games is a pragmatic argument about what is useful to forensics investigations.

However the story of use in the NSRL has more dimensions than this pragmatic argument about popularity and forensics initially suggests. The example of the Turkish intelligence service using the RDS to identify "terrorists" by locating the ByLock app on suspects devices is an example of *misuse*. NSRL team members recognized that their web collecting activities could be used by actors that did not share their values. The openness of the RDS data on the web confers the desired quality of neutrality that is so important for NIST as a standards body. But that same openness can lead to the RDS being used by a variety of actors whose interests do not align with the values of the NSRL and NIST. While the ByLock app did not in the end appear in the RDS, the dual-use nature of the NSRL was something that team members grappled with just below the surface.

Then there is the case of the Cabrinety Archive, which seemed to offer so little value in terms of the NSRL's forensic mission to collect popular software for the RDS. At the time of ingest at NIST most of the software titles were 25 years old. Four years after they were acquired the Cabrinety titles still had not been added to the RDS, in muted recognition that they would not actually be useful to forensics investigators. But this example of disuse is a hint that other uses and other users are at play, which opens to a view of the discursive work that the NIST-Stanford

collaboration is performing. The Cabrinety "completes" the archive by providing historical materials. It builds technical competencies inside NIST and at Stanford, and creates new networks of knowledge/power between defense, law enforcement, academia and cultural heritage. The values that drive appraisal have many sources, some explicitly celebrated for all, and some held close like stories for themselves.

In this analysis I have been drawing on the work of feminist scholar Sara Ahmed whose book *What's the Use* explores the "uses of use" and the "strange temporalities of use" (Ahmed, 2019). In recalling Foucault's project to study the "conduct of conduct" Ahmed adds a critical lens to the familiar utilitarian explanations of use which see *disuse* and *misuse* in purely negative terms–as things to be avoided and optimized away. Ahmed's artful repetition of photographs through her book highlights how a single scene, such as a postbox that is being used by a nesting bird, can be used to describe coexisting, overlapping use, disuse and misuse. More importantly perhaps she highlights how *queer uses* can function as a form of resistance, in opposition to dominant uses which are valorized.

Appraisal in web archives is not unlike conventional archives in that they are assembled for a particular use or uses. But the archive's architecture can generously encourage, or cynically inhibit, other uses, especially when the lifetime of records is drawn out in time. The "strange temporalities of use" are made possible by this extension in time. Ahmed's idea of *queer uses* resonated with me because it echoes Annemarie Mol idea of *ontological multiplicity*, that *practices* define different ontological realities which can coexist. All this is to say that multiple appraisal strategies can hang together, sometimes divergent, sometimes coherent, sometimes explicit, sometimes implicit, as heuristics of value, at a given time and (especially) over time.

6.3.2. Legibility

The other red thread that runs through my findings from the NSRL is the relation that holds between appraisal and *legibility*. I borrow this term *legibility* from anthropologist and political theorist James C. Scott whose excavation of modernist forms of seeing, and their failures, in *Seeing Like a State* highlights how important measurement is to what he calls high modernist programs. Here he describes how he was taken with this immense project while studying the attempts by the state to control nomadic peoples:

> How did the state gradually get a handle on its subjects and their environments? Suddenly, processes as disparate as the creation of permanent last names, the standardization of weights and measures, the establishment of cadastral surveys and population registers, the invention of freehold tenure, the standardization of language and legal discourse, the design of cities, and the organization of transportation seemed comprehensible as attempts at legibility and simplification. In each case, officials too exceptionally complex, illegible, and local social practices, such as land tenure customs or naming customs, and created a standard grid whereby it could be centrally recorded and monitored. (Scott, 1998, p. 2).

Given the findings I have detailed above it does not require a great leap of the imagination to recognize NIST's production of the NSRL as a legibility project. Indeed, their website clearly states the purpose of the NSRL:

> The National Software Reference Library (NSRL) is designed to collect software from various sources and incorporate file profiles computed from this software into a Reference Data Set (RDS) of information. The RDS can be used by law enforcement, government, and industry orga-

> nizations to review files on a computer by matching file profiles in the RDS. This will help alleviate much of the effort involved in determining which files are important as evidence on computers or file systems that have been seized as part of criminal investigations. (NIST, 2020a)

As the NSRL has shifted to collecting software from the web instead of physical media, it has needed to adapt its methods of appraisal to accommodate that new delivery mechanism. The use cases involving the Google Play Store and Steam illustrate how theories of popularity are baked into software, and how these automated appraisal processes can conflict with human agency as the legibility of what has been collected is lost.

The multiple generations of the *Juggernaut* database schema and its manifestation in different database systems speak to a history of legibility strategies that have evolved over time. The addition of keys and indexes while refactoring the database alter the performance profile of NSRL processes while also coercing users to interact with it specific ways. The files that make up *The Corpus* accumulate as new software is acquired; but they also open up to reveal new files as new capabilities to unpack container file formats are developed. New fixity values are recorded in the database, and in the RDS, as new fixity algorithms are studied, tested and deployed. Software is increasingly *network-contingent*, in that it can change its file based manifestation based on user interaction, and participation in a network of services. This fluidity prompts the NSRL to *disk-print* high value pieces of software in order to get the richest representation possible of the files that comprise a piece of software. The NSRL is a legibility project that is constantly under revision based on the types of software that are being acquired, and their understanding of how software packaging works, and the state of the art in hashing algorithms.

In a cognitive shift away from archival descriptive practices that are performed dur-

ing the accession of new materials, information studies scholar Elizabeth Yakel suggests that focusing on *representations* that evolve over time is more productive:

> Each successive representation and representational system builds on its predecessors, recovering what was judged valuable in a given temporal and cultural context, incorporating or discarding what was deemed essential or not, respectively (Yakel, 2003).

The connection that Yakel makes here between representation (legibility) and the value judgment (appraisal) is a key insight here. The desired forms of legibility in the NSRL shape the types of software that are collected.

The topic of legibility raises the questions of what is being made legible, and for whom. The answer that immediately suggests itself is that the production of the RDS makes unique files on a computer storage systems legible so that forensic investigators can ignore them, and focus on the unique files. Thus the RDS functions as a type of *finding aid* for the software in the NSRL. But this is no ordinary finding aid because it allows the NSRL to be viewed as a type of *negative-archive*, or means for inverting attention rather than attracting it.

Viewed as an anti-archive helps to bring into focus how software is being made legible by the NSRL not to people, but to machines. People don't read the RDS like they would a traditional finding aid. The RDS is an algorithmically generated finding aid that is designed to be read by other algorithmic systems. The assemblage of forensics tools and investigators that are produced, tested and promulgated by the NSRL's sister project the Computer Forensics Tool Testing (CFTT) program provide a type of *data vision* that emerges out of the interaction between specialists, tools and data:

> data vision: the ability to organize and manipulate the world with data and algorithms, while simultaneously mastering forms of discretion

> around why, how, and when to apply and improvise around established methods and tools in the wake of empirical diversity (Passi & Jackson, 2017).

This sociotechnical interaction that situates human improvisation and algorithmic processes in partnership to achieve some measure of *data vision* is similar to the idea of legibility or vision proposed by sociologist Janet Vertesi in her book *Seeing Like a Rover*. Here Vertesi compares her approach to Scott's when studying the imaging practices of the Mars Rover:

> Like seeing like a state, seeing like a Rover also requires mutual entanglement of ordered vision and institutional agency. The resulting images enroll multiple observers in complex social relations, but these relations are oriented toward consensus, not authoritarian control. Even though rover images are disseminated by a government authority (NASA), observing behind the scenes reveals how images are enrolled in producing a collectivist visual experience: built from the bottom up, shared across the mission team, naturalizing knowledge production on Mars, and reinforcing social orderings on Earth. (Vertesi, 2015, p. 16)

Vertesi approach suggests a generous, and arguably less critical, view on the activities of the NSRL. The entanglement of algorithmic processes of seeing with social practices of knowledge production is arguably a type of legibility that the NSRL is pursuing when it creates the RDS and as it considers the positive-archive it has created: all the disk images and bags of software that have been collected over the years, and their potential for circulation and study that we see hinted at in the Stanford-NIST collaboration. However while the RDS as a dataset is open, outside the Steering Committee there is very little transparency around when and what software is being acquired by the NSRL, what is being processed and hashed and un-

packed. In short *The Numbers*, or the rate of change in processing in the NSRL are not presented outside of verbal performances in closed meetings, and aggregated statistics in RDS releases. If the NSRL is to grow beyond its use as a negative-archive strictly for forensics purposes it must open itself up further, and encourage new knowledge production practices and social interactions.

6.3.3. Governmentality

The factors of *use* and *legibility* in the NSRL are mutually constitutive, since: specific uses of the archived software in the NSRL require that records be made legible in particular ways; and the ways that that records can be made legible shape what types of uses are possible. Attending to this interaction between use and legibility helps perforate the claim to neutrality that's implicit in thinking of the NSRL as "one of the largest software libraries in the world" (NIST, 2018). Looking below the surface of the NSRL's forensics use case, and observing the wide range of use/legibility processes at work in the NSRL allows us to see that the web is being archived and appraised.

But as we saw in the NIST-Stanford collaboration, use and legibility also work hand in hand to discipline the field of digital preservation. The processes and workflows developed at NIST for forensics investigations are put to work to migrate the Cabrinety collection from its original media into forensic disk images that are then sent to Stanford. In the process NIST gained experience with older physical media containers while building its own historical collection. Similarly Stanford was able to bootstrap access to its software archive by ingesting the forensic disk images received from NIST into their digital preservation repository. These same forensic disk images go on to participate in newly developed software distribution networks that generate new possibilities for providing historical research, which make possible new interpretations of copyright law.

At the risk of overcomplicating this schema for appraisal in web archives I think it is useful to see this recursive interchange between use and legibility as an example of what Michel Foucault calls *governmentality*. You may recall from Chapter 2 that governmentality is not specifically concerned with the arts of government in the conventional sense (the state), instead it's interested in the *conduct of conduct*, or all the activities that shape how people behave. Governmentality is especially interested in the specific systems of rationality that are used to support these activities. My findings suggested that NIST's juridical digital forensics tools found expression not only in courts of law but as digital preservation practices for cultural heritage organizations. The practices of disk imaging and fixity analysis are tools for measuring people through their hard drives. NIST's practices of appraisal of the web, their decisions of what to collect, and what not to collect, are an enactment of this forensic apparatus.

Foucault is perhaps best known for his critical analysis of power as domination, such as his historical excavation of how surveillance architectures designed for prisons find their way into the design of factories and schools (Foucault, 2012). He developed the idea of governmentality later in life, to provide an analysis of power that is relational and *productive*, which traces how systems of measurement and calculation are used to make populations and subjects legible, in order to further ideas about life and health, or what he also calls *biopower*. Niklas Rose' uses Foucault's idea of governmentality to trace the development of the sociotechnical theory from the experience of managing the military and society during World War 2 into present day disciplines of psychology and sociology:

> Michel Foucault argued that the disciplines "make" individuals by means of some rather simple technical procedures. On the parade ground, in the factory, in the school and in the hospital, people were

> gathered together en masse, but by this very fact they could be observed as entities both similar to and different from one another. These institutions function in certain respects like telescopes, microscopes, or other scientific instruments: they established a regime of visibility in which the observed was distributed within a single common plane of sight. Second, these institutions operated according to a regulation of detail. These regulations, and the evaluation of conduct, manners, and so forth entailed by them, established a grid of codeability of person attributes. They act as norms, enabling the previously aleatory and unpredictable complexities of human conduct to be charted and judged in terms of conformity and deviation, to be coded and compared, ranked and measured. [Rose (1999); pp. 135-136]

This analysis here can serve as a template for viewing the operations of appraisal in the NSRL. The "grid of codeability" directly speaks to the RDS and its use in forensics tools. For Foucault and Rose who came after, power is not something that is localized wihin a particular entity like the NSRL or NIST, it is distributed throughout society–it is relational involving many different subjects as diverse as historians, archivists, game publishers, standards bodies, law enforcement and child pornographers.

7. A Conclusion

The last three chapters have described how appraisal practices in web archives are a rich and multifaceted practice. In Chapter 4 we saw how appraisal work is fundamentally sociotechnical. In appraising content archival practitioners must direct the use of automated tools that do the work of selecting and retrieving content from the web. But they also engage in significant improvisational work to make these tools, which are optimized for algorithmic data collection, fit their ideas of appraisal as a legible and collaborative activity. In Chapter 5 we saw how the architecture of the web disrupts the relationship of trust between records creators and archives, which is a central feature of appraisal. However, archivists continue to work in a vibrant community of practice, where this rupture in trust can be mended with more attention to the accountability and positionality of web archives. Finally, in Chapter 6, I unpacked how the values inherent in appraisal are dependent on the twinned phenomena of use and legibility, which operate as an expression of governmentality.

I set out in the beginning using method-triangulation to generate a thick description of appraisal practices in web archives. In that regard I think this project has been successful. But the goal in creating this description was to *zoom in* from what is talked about (themes), to how it is talked about (discourse), to what actually happens (ethnography), in order to zoom back out again, and draw some conclusions. The problem with thick description is that the resulting description is, well, thick: it's dense, knotted and difficult to untangle. This makes zooming out from all the details difficult. Despite the difficulty, I will conclude with few observations of how these three studies hang together, and what they say about appraisal in web archives.

One way of synthesizing the findings in these studies is to reflect on the key controversies that I encountered while conducting them. By controversies I don't mean specific events that were found in my data or findings. Instead I'm interested in

my own conceptual crises; where what I learned or experienced challenged hidden assumptions of what appraisal in web archives is. Were there hidden givens, or unstated axioms of appraisal that I had at the beginning of my project which my findings compelled me to reevaluate? Fortunately the answer is yes, and that they fell roughly into the categories: time, ontology, and use.

Recall the focus on *seed lists* in Chapter 4. I treated seed lists as singular artifacts that recorded appraisal decisions about web content. Seed lists are instructions for software to collect particular websites. In my interviews I chose to ask participants about how URLs were added to their seed lists thinking that this would give me access to their thinking about appraisal decisions. Similarly, the analysis in Chapter 5 was geared towards drilling down into the moments of *archivalization* in which the searchlight of appraisal was manifested in decisions to archive one thing instead of another. Somewhere in the middle of my NIST field study I realized that there was no single moment in which records were appraised. Certainly, initial decisions were made to collect some software instead of others. But the criteria used was constantly under revision due to environmental concerns (storage, bandwidth, auditing concerns). Also, the records that were acquired, the many computer files that comprise software, were not as fixed as I initially thought, and were appraised again, and again as files gave rise to other files as container file formats were recognized and unpacked. The descriptions of the files, the RDS, which was the purest expression of the archive, was itself changing over time, and sometimes even its semantics (the fixity algorithms) were changed, which required reprocessing again and again. Thus, appraisal, or the expression of a record's value is not evaluated once at a particular phase in a life-cycle of a web archive, instead it is spread across many archival processes, which repeat and evolve.

The second given that I failed to fully register at the beginning of my project was

the ontology of web archives, or what constitutes a web archive. I started out my project thinking I would study the places on the web that get called "web archives", such as the Internet Archive. Web archives are commonly thought to be specialized infrastructures that collect representations of web resources (web pages), so that they can be "played back" later. Web archives may use different software to achieve these ends, but to some degree they all crawl the web, save WARC data, and index it, in order to play back what specific URLs looked like at a particular time. In other words web archives attempt to recreate the experience a person has when looking at a web page in their browser in a particular moment.

My field study with NIST came together once I recognized that even though it did not fit the usual architectural mold of a web archive, the NSRL was deciding to collect things from the web, preserve them, and provide access to them, and thus, it is a web archive (at least in part). The NSRL was not interested in crawling the web in order to play back what web pages looked like at a given time. Instead it was interested in presenting a machinic view into what software is being made available on the web (the RDS). This realization opened my eyes to the possibility that there are in fact many *shapes* of web archives to consider in archival studies. Artificially limiting the study of web archives to one specific technical shape significantly truncates the scope of phenomena we need to be considering when we study and talk about web archives.

The third hidden axiom that I came to recognize during my research concerns the concept of the archival record and its relation to appraisal value. Without paying it much thought I assumed that appraisal value was a property, or attribute of the record. Based on the research literature I thought archivists, if they were lucky, would have an appraisal policy, collection development policy, or at least some shared notion of what records the archive collected and why. I thought that new records were

evaluated using this criteria, and a decision was made to either accession or discard the records. I've already discussed how I was mistaken about appraisal decisions happening in a singular moment. However I was also mistaken about records having a specific value as an attribute. The value of records acquired from the web by the NSRL were bound up in how the records were to be used. An essentialist view of appraisal, where value is a property of records, is replaced by a relational one, where records obtain value based on their use.

Initially this shift in attention to use seemed like a familiar pragmatist line of argument about value, where the ends justify the means. But further investigation, and reflection with Ahmed (2019) revealed that some records were acquired even though they had no obvious use (disuse). Some records were in a superposition of use and misuse (ByLock and SWID). Use is fluid, multiple, contradictory, and subject to resistance. Perhaps it is more accurate to say that appraisal value is inherently tied to *practice* to avoid the utilitarian interpretation of the word "use".

The consideration of the use of records has generally been the third rail of appraisal theory. For example in his 12 Principles of Appraisal, Richard Cox highlighted:

> The archival appraisal selection criteria should rest not on unpredictable future research practices and trends but upon the more predictable sense of determining what are the salient and important features of contemporary institutions and society. (Cox, 1994, p. 24)

Here the archivist is forbidden to think about the use of records, because it is difficult to predict the future. Instead the archivist is asked to document society, as if this is somehow easier. But the archivist does have access to the past, and the present–and does have some ideas about how records have been used, and can be used in the present. In principle the *use of use* and its application to archival appraisal is not unlike the concept of functional analysis, which is at the core of appraisal theories

like macro-appraisal (Cook, 2004) and documentation strategies (Samuels, 1991). Here is how Cook describes this shift from the record to *function*:

> … macro-appraisal shifts the initial and major focus of appraisal from the record–and any research characteristics or research values it may contain–to the functional context in which the record is created–its conceptual, virtual or functional provenance. Using knowledge gained by an institutional functional analysis, including an analysis of the interaction of function and structure, of organisational cultural dynamics, of record-keeping systems and of citizen/client involvement and interaction with the institution or function, the main appraisal questions for the archivist become, first, what functions and activities of the creator should be documented (rather than what documentation should be kept?) and, secondly, who—in articulating and implementing the key functions, programmes and transactions of the institution—would have had cause and the primary responsibility to create a document, what type of document would it be, and with whom would that corporate person interact in either its creation or its later operational use? These two questions suggest a third: which record creators or 'functions' (rather than which records) are the most important? (Cook, 2004, pp. 9–10)

Cook acknowledges elsewhere that his definition of macro-appraisal (and its use of functional analysis) relies heavily upon the citizen/state relation, but that the principle should hold for relations that exist between company/customer, university/student, hospital/patient, union/member, church/parishoner, etc. For Cook the focus on the *functions* inherent in these relations helps to put the question of records' value into the background. This move is especially important in order to factor out any guesswork about potential research value. However, notice how Cook positions

the archivist on the outside looking in at the functions, or as I have called them, uses? In my field study with NIST participants were making appraisal decisions because they were in the process of building an archive for use, these records had a particular function. Feminist and STS scholar would call this an example of the *God Trick*:

> I would like to insist on the embodied nature of all vision and so reclaim the sensory system that has been used to signify a leap out of the marked body and into a conquering gaze from nowhere. This is the gaze that mythically inscribes all the marked bodies, that makes the unmarked category claim the power to see and not be seen, to represent while escaping representation. (Haraway, 1988, p. 581).

Cook and Samuels position the archivist as an analyst/judge one step removed from the site of record creation. As evidenced in the NSRL-Stanford collaboration to archive the Cabrinety collection, even when use isn't completely understood, record creation is always an expression of power/knowledge. If some story of use, disuse or misuse isn't visible something is missing from the picture. At least Schellenberg owned that records are (in part) collected for use by researchers (Schellenberg, 1956). Research isn't the only use for archival records, but it a significant one nonetheless. More importantly *use* recognizes that records are created for particular purposes. The value of records lies in an analysis of their genealogy of use, disuse, misuse, and queer use. Where we can imagine queer use as unexpected value that is discovered in research.

> The project of queering use does not aim to create distance from use but to inhabit use all the more. We might respond to the problem of instrumentalism not by rejecting the idea of useful knowledge but by calling for knowledge that is useful to others, with this "to" being an opening, an invitation, a connection. (Ahmed, 2019, p. 222).

At the risk of naturalizing these genealogies of use, thereby erasing the role of archival agency, one way to visualize these genealogies of use is as the concentric rings of a tree trunk which spiral outwards, just as data is relayed from one context to another (Janée et al., 2009).

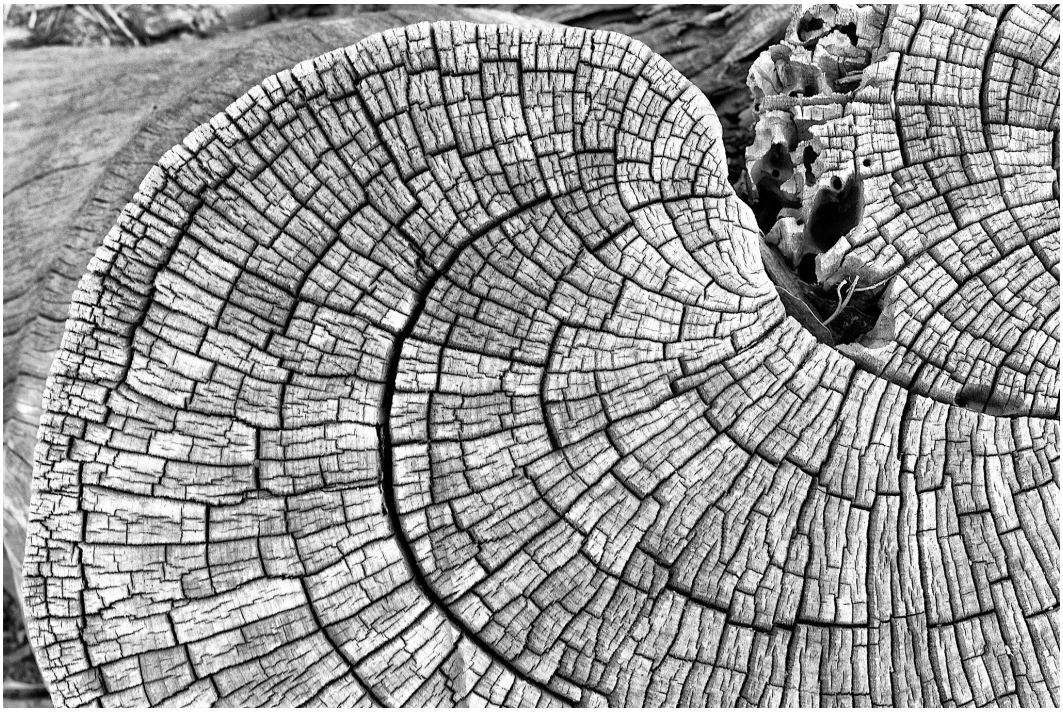

Figure 24: 2013/465/24 Time Curves by Alan Levine

Was the bark created to record the age of the tree, or to store information about the amount of carbon-dioxide in the atmosphere, or to indicate how much rainfall there was that year? No. The bark's initial use is to protect against damage from parasites, animals, diseases, dehydration, and fire. As the tree ages the bark supports the tree. Any secondary use of the bark over time, such as to measure age, derives from an understanding of the bark's initial use.

While provenance is normally understood in terms of ownership, this genealogy of use is at the heart of what provenance is concerned with. Below in Figure 25 is a diagram of these concentric rings of use, where the initial use of the records is labeled as U₀ and subsequent uses as U_n, U_{n+1}.

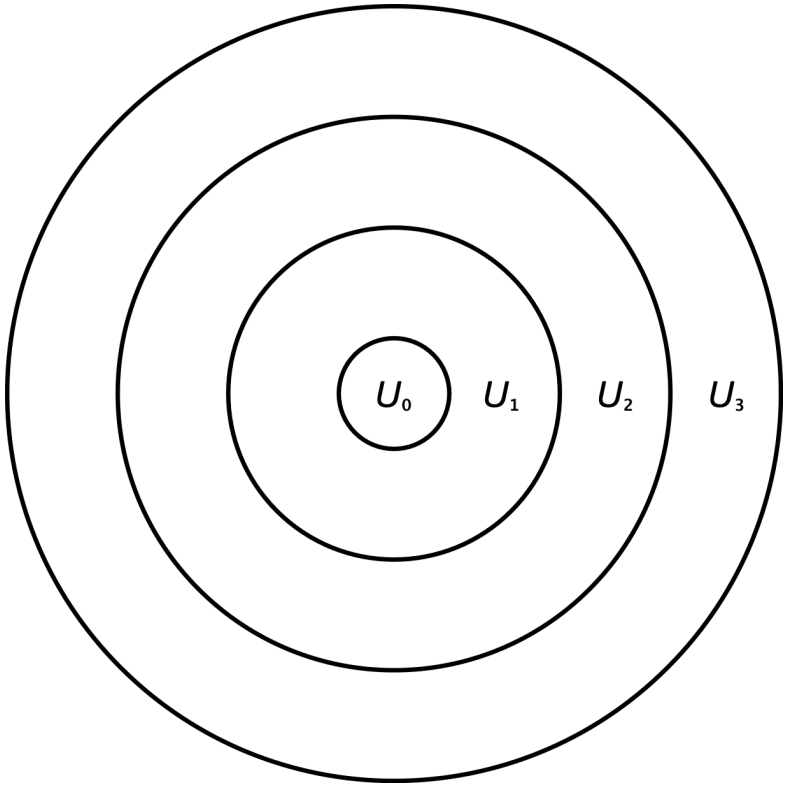

Figure 25: Concentric Rings of Use

Failing to attend to the use of records lets us believe in the fantasy of their singular use, as if they are always and only the evidence of a particular thing. But allowing use into the picture allows us to appreciate the full set of contingent relations and value propositions that web archives participate in. Perhaps it is easier to recognize these relations when considering digital records, such as web archives, because as data they can easily copy and transport themselves into new contexts. Figures 26 and 27 extend the single series of use by illustrating how a series can be doubled where multiple uses are present at one time: `U<sub>0</sub>` forks into `U<sub>1</sub>` and `V<sub>1</sub>`. It is in these genealogies of use that we see the most complete expression of the values that web archives entail. Archivists who divorce the collection of records from the use of those records will be forever chasing their own tail when trying to understand the value of records. Recognizing, critiquing and celebrating the use of records is where we encounter the value of web archives.

Figure 26: Age in Double Figures? by Garry Knight

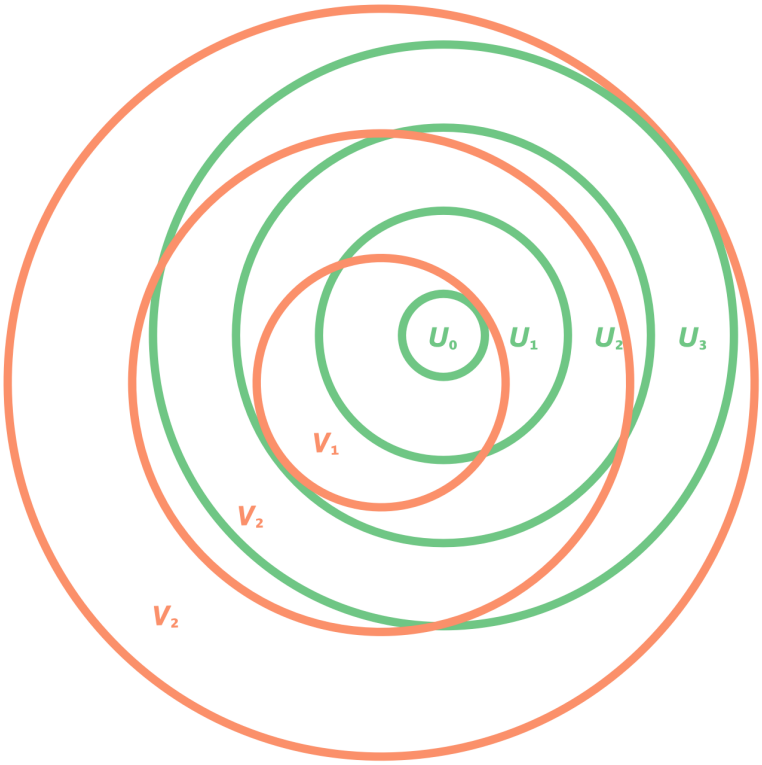

Figure 27: Multiple Concentric Rings of Use

Dismantling these three hidden axioms (time, ontology and use) of appraisal in web archives clears the way for several areas of values based research for web archives and archival studies more generally. Archival studies researchers have a much larger and more variegated landscape to observe and analyze once the architecture of the web archive is freed from the notion that it must fit a particular architectural shape, or technical mold. For example what archival practices do individuals enact with their own social media data? How do the "archives" offered by web platforms operationalize a view of what an archive is, and how it should be used? By what processes are websites migrated forwards in time as the contents of one content management system are migrated to another? How do the archives of web content collected in one context transfer as assets into another context during corporate acquisitions and mergers? How do distributed web technologies such as IPFS (Protocol Labs, 2020) or the partnership between CloudFlare and the Internet Archive (Graham, n.d.) enact a particular archival view of the web? And as machine learning models are generated from large accumulations of web content how do we adequately document these web archives so that the biases inherent in the models can be interpreted?

An example can help make this last question a bit more concrete. OpenAI's GPT-3 deep learning language model is trained using text collected from the web by the CommonCrawl project (Weinberg & Zimmermann, n.d.). What parts of the web are being collected by CommonCrawl? How do we understand the values of these vast collections of data and language models as web archives that are being created for particular uses? Casting these research topics as questions for archival studies research and not simply the concerns of machine learning or human computer interaction (HCI) research is critical for understanding them not simply in terms of innovation but as value driven efforts that extend over time.

But I think this dissertation's consideration of archival appraisal in the context of

web provides some insight into problems that have preoccupied archival studies more generally. In a 2013 issue of the *American Archivist* Mark Greene, Randall Jimerson and Michelle Caswell engaged in a heated debate about the place of social justice in archival studies. To overly simplify a complex set of arguments made by all three, Greene contended that a social justice agenda in archives weakened the profession, because once activated, it limits the archivist's ability to collect (e.g. from the opposing side of an issue). Jimerson responded saying that he did not mean to suggest that all archivists should assume the mantle of social justice, and that all archivists needed to exercise their abilities and conscience–the world needs more than one archives. Caswell responded primarily to Greene by saying that an inadequate engagement with what constitutes *social justice* by caricaturing a small number of archival studies scholars undermined any argument he was attempting to make. However, at the center of Greene's provocation was an interesting paradox, which he presented by way of Ketelaar (2005), which has other parallels:

> The records created and used by German and Dutch agencies during the Second World War to account for the looting of Jewish assets continued to be used, after the war, by German and Dutch agencies in the processes of restitution and reparation. The same record was activated by different societal powers, for different purposes and for different audiences again and again, as it is today activated in the search for looted and lost works of art and other Holocaust assets. The looting and the registration of the looted property were, of course, an appalling event, but it was through the subsequent uses of the record that the primary registration became really a record of a traumatic experience. (Ketelaar, 2005, p. 296).

How can archivists make appraisal decisions that reflect an alignment with social

justice? If a German archivist exercises their agency by distorting or refusing to manage the records of looting then their subsequent use for returning the stolen items to their rightful owners would not have been possible. However, if they chose to create and manage these records to enable the theft then they are a willing participant in an ethically bankrupt activity. In some ways this is a case of 20/20 hindsight. But I think it's useful to accept the paradox of these records as a philosophical problem.

The philosopher Ludwig Wittgenstein famously saw paradoxes as linguistic misunderstandings which dissolve with the therapy of philosophy.

> It is not our aim to refine or complete the system of rules for the use of our words in unheard-of ways. For the clarity that we are aiming at is indeed *complete* clarity. But this simply means that the philosophical problems should *completely* disappear. The real discovery is the one that makes me capable of stopping when I want to. (Wittgenstein, 1953, p. 133)

Insights that allow problems to disappear rather than be solved are what philosophy is all about. Later in life Wittgenstein proposed that words weren't defined in terms of their reference to objects in the world, but in terms of their use in human activity. Records are similar in that they have no value outside of considerations of use. There may be many uses, some uses may be hidden, some uses may conspire against each other, some uses may as yet be unknown, but it's only in recognizing and understanding their use that a record's value can be understood, and acted upon.

8. Appendixes

A. Study 1 Documents

A.1. Consent Form

You are invited to be in a research study that explores the selection of Web content for preservation in an archive. You were selected as a possible participant because you have some expertise in either the selection of Web content for archival processing or the design of tools to assist in the archiving of Web content. I hope to interview you on these subjects.

Interviewing will possibly occur with real-time computer and code referencing, this is the considered the "observation" element of this study. Please feel free to ask me any questions before participating. Being in the study is voluntary and you are free to stop at any time. Refusing to be in the study or stopping study activity will involve no penalty or loss of benefits to which you are otherwise entitled.

The purpose of this document is to give you the information you will need to help you decide whether to be in the study or not. Please read the form carefully. You may ask questions about the purpose of the research, what I would ask you to do, the possible risks and benefits, your rights as a volunteer, and anything else about the research or this form that is not clear. When I have answered all your questions, you can decide if you want to be in the study or not. This process is called "informed consent." I will give you a copy of this form for your records.

Background This study is being conducted by Ed Summers in the College of Information Studies at the University of Maryland. Its purpose is to explore the current processes and tools used to select content from the Web for archiving. If you agree to be in this study, I ask your permission to a conduct face to face semi-

structured interview.

Your decision on whether to participate will not affect your current or future relations with the University of Maryland or your current employer, and you may withdraw at any time without affecting those relationships. The face to face interview usually takes no more than 1 hour, and there is no compensation for participating in the face to face interview. The observation component of this study may be part of the interview and is included in this 1 hour time period. I envision this as an active interview that possibly involves looking at Web content together, talking about Web archiving work, and looking online for examples to discuss.

The audio/video recordings will be transcribed and then coded to identify patterns and strategies for selecting Web content for archiving.

Risks There is no physical or medical component to this research, and there is no risk of physical injury. The identities and organizational membership of the interviewees will not be revealed in our published findings, and a pseudonym will be used.

Benefits There may be no benefits to you personally for participating in the current research; however, there may be some professional and societal benefits. This study will help the archival community better understand the processes by which Web content is selected for an archive. Currently very little research into this phase of the work of Web archiving has been done. The hope is that the findings of this study will help influence the design of tools that assist archivists in their work.

Confidentiality Confidentiality will be maintained by a) aggregating information and b) by assigning pseudonyms. I have an ethical and legal obligation to protect confidential information used or obtained in the course of research and all policies

on confidentiality apply equally to data stored both in the computer and on paper records. Any non-disclosure agreements to which you are a party will be respected and maintained by the security of aggregation and pseudonyms.

Confidentiality will be maintained by aggregating information in tables and graphs that describe broad trends in attitudes and demographics across the population. Composite descriptive sketches will not refer to the data collected from named individuals.

Confidentiality will be maintained by assigning you a pseudonym. Any taped conversations, paper notes, or other research materials associated with our exchanges will be identified with the pseudonym. The only code sheet identifying you with your pseudonym will be kept in locked storage over a mile away from the research materials. This code sheet will be destroyed at the end of the study, or by the end of 2016.

If any published material is going to include a quotation from the transcript I will notify you via email and request your approval. If I don't hear anything back in two weeks I will resend the notification. If there is no response I will publish the quotation. I will respect your wishes to either not include the quotation, or to provide clarification.

Contacts You will be given a copy of this form to keep for your records, and you may ask any questions you have now. If you have questions later, you may contact me by telephone at +1 (240) 478-7086 or by email at edsu@umd.edu.

If you have questions about your rights as a research participant or wish to report a research-related injury, please contact:

University of Maryland College Park Institutional Review Board Office 1204 Marie Mount Hall College Park, Maryland, 20742 E-mail: irb@umd.edu Telephone: 301-

405-0678

This research has been reviewed according to the University of Maryland, College Park IRB procedures for research involving human subjects.

Statement of Consent Your participation in this study indicates that you are at least 18 years of age; you have read this consent form or have had it read to you; your questions have been answered to your satisfaction and you voluntarily agree to participate in this research study.

A.2. Recruitment Letter

Dear {Name}

I am a researcher at the University of Maryland, and I am investigating the appraisal process in Web archives. I am interested in interviewing you regarding your expertise in this matter and attached to this email is the formal interview request. If you agree the interview will last no longer than 1 hour. Please feel free to contact me with any questions you might have.

You were selected as a possible participant because you have some expertise in either selecting Web content for an archive, or in building tools/services that facilitate the selection process.

I will be conducting the interviews via Skype and recording the audio and video. This is considered the "observation" element of this study. In the analysis phase of the study I will personally be creating transcripts, and performing open coding on them for theory building about the Web appraisal process. Your name and your institution will not be mentioned in any published material from these transcripts.

I hope you volunteer to take part in this study. Please contact me if you are interested in participating or would like to ask me any questions about it. If you do not contact me, I will follow up with you once and then assume that not hearing from you means you are not interested in participating. Please note that communication sent via e-mail cannot be guaranteed confidential.

Please find attached an information statement with additional information on the interview procedure.

Sincerely,

Ed Summers

A.3. Interview Protocol

First, I want to thank you for participating in this interview today. I know you are busy and I really appreciate you taking the time to help me. Before we get started I thought I'd start by telling you a little bit about the study and what you can expect from this interview.

Just to review, the title of the study is Investigating Appraisal in Web Archives. As you may know there are many initiatives to archive parts of the Web. These can be found in libraries, archives, museums, businesses and government. Often these organizations have articulated collection development policies to help guide what Web content is collected. However the actual process for discovering websites and content that is relevant for a Web archive isn't well known. The Web is a big place, and even at the organizational level it can be difficult to know what needs to be collected and when.

In this study I'd like to talk to you about this process of selection or appraisal. The hope is that a better understanding of the decisions and mechanics of how archivists select content will help inform the design of new tools to assist archivists in their work. I'm really interested in the nitty gritty practicalities about how websites are added to an archive. I do have a short list of questions but this is going is a semi-structured interview because I would like the conversation to evolve organically. I'm pretty sure I don't have all the right questions!

I'm anticipating that the interview will take anywhere from half an hour to an hour. Feel free to say you don't know the answer to a questions, or you'd rather not answer it, and you can end the interview at any time. Please ask me to clarify any questions that aren't clear. I am recording the audio and video from the interview, which I will be transcribing and then doing content analysis to (hopefully) find patterns and themes. Your interview will be kept confidential and destroyed at the completion

of my study. I will not mention you or your institution by name in my study except using a pseudonym. If I plan on quoting you at all in published material I will modify the excerpt to reduce the likelihood of identification. I will also contact you and ask for your permission to use the segment.

Do you have any questions about what I just covered or anything else?

Great, well lets get started then.

1. Could you tell me a little bit about your role at XXX? (Alt: how long have you been there? What are your responsibilities?)
2. Can you describe your work environment? (Alt: How many people do you immediately work with, what are their roles?)
3. Have you ever selected Web content to be archived? (Alt: do you do this on a regular basis?)
4. Try to recall a time when you selected content for archiving. What was it? Can you remember how you found it?
5. Can you think of another example?
6. Can you think of any other examples that seem different from these two?
7. What criteria do you think go into deciding whether a particular Web page or website is worth collecting? Can you give me examples?
8. How does the presence of the content in other Web archives factor into your decision if at all?
9. Do you consider whether it is important to collect the same web resource over time?
10. Do you record any information about why a particular website was selected?

Do you think that could be useful?

11. How does the perceived size of a website factor into your decision to archive content?
12. Do you search for content to archive? How do you do it?
13. Do you consult with your peers when locating Web content? How does that work?
14. Do you talk to peers at other institutions or organizations when identifying web content?
15. Do you ever interact with the content owner/provider when selecting content for archiving?
16. Do people ever try to donate Web content? Can you describe an example of that?
17. Do researchers ever request that you archive particular Web content? How did that happen?
18. Have there been any requests to access archived content yet by internal or external people? How did that go?
19. Could an inventory of your Web archives contents be made available for a follow up study?
20. Does your organization have a collection development policy when it comes to collecting Web content? What kind of guidance does it provide? (Alt: are there any agreed on criteria for what Web content to archive?)
21. Is that policy available to the public? Can I get a copy?

B. Study 2 Documents

B.1. Consent Form

Project Title Enacting Appraisal: Investigating the sociotechnical factors of software selection in the NSRL.

Purpose of the Study This research is being conducted by d Summers at the University of Maryland, College Park. I am inviting you to participate in this research project because you have worked to help build and/or sustain the National Software Reference Library. The purpose of this research project is to better understand how decisions are made about how to build digital collections, particularly when they involve obtaining content from the web.

Procedures The procedure involves an unstructured interview with you which will be audio recorded and transcribed. Interviews will take between 60 and 90 minutes. If you wish you can choose a pseudonym that will be used in place of your name in all written materials. Content analysis performed on the transcripts will be used to derive emergent themes and issues that speak to the study's research question. There are no direct benefits to participants.

Potential Risks There is no more than minimal risk associated with participating in this study. However, every potential subject will have the option to refrain from participation. Additionally, all participants will be asked if they would like me to use a pseudonym instead of their name in order to protect their identities and minimize risk.

Potential Benefits There are no direct benefits to you. However, the goal of this study is to understand how content selection processes and technical infrastructures

have coevolved in the NSRL. This may or may not be of interest to your work in or with the NSRL.

Confidentiality Any potential loss of confidentiality will be minimized by storing data in encrypted form in a private DropBox folder which will only be accessible by researchers. If I write a report or article about this research project, your identity will be protected to the maximum extent possible. I will also notify you before hand if I plan to use any quotes from your interview to give you an opportunity to clarify, or if you would prefer me not to use the quote.

Your information may be shared with representatives of the University of Maryland, College Park or governmental authorities if you or someone else is in danger or if we are required to do so by law. If participants wish their interviews can be donated back to the NSRL as historical documents.

Right to Withdraw Your participation and Questions in this research is completely voluntary. You may choose not to take part at all. If you decide to participate in this research, you may stop participating at any time. If you decide not to participate in this study or if you stop participating at any time, you will not be penalized or lose any benefits to which you otherwise qualify.

If you decide to stop taking part in the study, if you have questions, concerns, or complaints, or if you need to report an injury related to the research, please contact the investigator:

Ed Summers

Maryland Institute for Technology in the Humanities
University of Maryland, College Park, 20742
edsu@umd.edu
301.405.8927

If you have questions about your rights as a research participant or wish to report a research related injury, please contact:

University of Maryland College Park
Institutional Review Board Office
1204 Marie Mount Hall
College Park,
Maryland, 20742
Email: irb@umd.edu
Telephone: 3014050678

For more information regarding participant rights, please visit: https://research.umd.edu/irbresearchparticip

B.2. Recruitment Letter

Dear _____

My name is Ed Summers, and I am a researcher from the College for Information Studies at the University of Maryland. I am spending a year working with the National Software Reference Library as part of NIST's Professional Research Experience Program (PREP). During this time I'm investigating the social and technical factors that help shape the construction of digital archives, in particular the content that they collect.

Given that the NSRL is a unique and long running example of a digital archive I was hoping to speak with you about your experience working in or with the NSRL. Hearing more about the types of work you do, and how you think about the activities of the NSRL would be extremely helpful to me as I conduct this research project.

All interview materials (transcripts and recordings) will be kept confidential. If you would rather not be named as a participant you can choose a pseudonym that will be used in all research materials. In addition if any quotations from the interview that are used in published materials they will be sent to you beforehand for clarification and approval. I expect the interview to last no longer than one hour. I hope we can find a time to talk.

Sincerely,

Ed Summers

B.3. Interview Protocol

1. How did you first come to work with the National Software Reference Library?

2. Can you describe what your usual working day is like? For example, what kinds of activities do you get up to, and who do you interact with the most, and has this changed over time?

3. What computer systems, applications or tools do you use most often in your work?

4. How is material being selected for the NSRL? What do you think

5. How has the NSRL Reference Data Set and the collection of software been used?

6. Is there anything you were expecting me to ask which I didn't?

9. References

Abreu, A., & Acker, A. (2013). Context and collection: A research agenda for small data. *IConference Proceedings*, 549–554.

Acker, A. (2015). Radical appraisal practices and the mobile forensic imaginary. *Acrhive Journal*, *5*. http://www.archivejournal.net/issue/5/archives-remixed/radical-appraisal-practices-and-the-mobile-forensic-imaginary/

Ahmed, S. (2019). *What's the use: On the uses of use*. Duke University Press.

Andersen, H. (2013). A website owner's practice guide to the wayback machine. *J. On Telecomm. & High Tech. L.*, *11*, 251. http://www.jthtl.org/content/articles/V11I1/JTHTLv11i1//_Andersen.PDF

Anderson, K. D. (2011). *Appraisal learning networks: How university archivists learn to appraise through social interaction* [PhD thesis]. University of California, Los Angeles.

Arnold, T., & Sampson, W. (2014). Preserving the voices of revolution: Examining the creation and preservation of a subject-centered collection of tweets from the eighteen days in egypt. *The American Archivist*, *77*(2), 510–533.

Aronson, J. D. (2017). Preserving human rights media for justice, accountability, and historical clarification. *Genocide Studies and Prevention: An International Journal*, *11*(1).

Aufderheide, P., Butler, B., Cox, K., & Jaszi, P. (2018). *Code of best practices in fair use for software preservation*. Association for Research Libraries.

Axon, S. (2019). *You can now download the source code for all infocom text adventure classics*. https://arstechnica.com/gaming/2019/04/you-can-now-download-the-source-code-for-all-infocom-text-adventure-classics/

Baeza-Yates, R., Marin, M., Castillo, C., & Rodriguez, A. (2005, May). Crawling a country: Better strategies than breadth-first for web page ordering. *Proceedings of the 14th International Conference on World Wide Web*.

Bailey, J. (2013). Disrespect des fonds: Rethinking arrangement and description in born-digital archives. *Archive Journal*, *3*.

Bandaru, K., & Patiejunas, K. (2015). *Under the hood: Facebook's cold storage system.* Facebook. https://code.fb.com/production-engineering/-under-the-hood-facebook-s-cold-storage-system-/

Banos, V., & Manolopoulos, Y. (2015). A quantitative approach to evaluate website archivability using the clear+ method. *International Journal on Digital Libraries*, 1–23. https://doi.org/10.1007/s00799-015-0144-4

Barth, A. (2011). *HTTP state management mechanism* (Nos. 6265). Internet Engineering Task Force. https://tools.ietf.org/html/rfc6265

Bastian, J. A. (2001). Taking custody, giving access: A postcustodial role for a new century. *Archivaria*, *53*, 76–94.

Bearman, D. (1989). Archival methods. *Archives and Museum Informatics*, *3*(1). http://www.archimuse.com/publishing/archival//_methods/

Ben-David, A., & Amram, A. (2018). The Internet Archive and the socio-technical construction of historical facts. *Internet Histories*, *2*(1-2), 179–201.

Benjamin, W. (1999). The work of art in the age of mechanical reproduction. In *Illuminations: Essays and reflections* (pp. 211–244). Pimlico.

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom*. Yale University Press.

Berners-Lee, T., & Fischetti, M. (2000). *Weaving the web: The original design and*

ultimate destiny of the world wide web by its inventor. Harper.

Boland, T., & Fisher, G. (2000). *Selection of hashing algorithms*. https://web.archive.org/web/20041015180613/http://www.nsrl.nist.gov/documents/hash-selection.pdf

Boles, F., & Young, J. (1985). Exploring the black box: The appraisal of university administrative records. *The American Archivist*, *48*(2), 121–140.

Bolter, J. D. (2016). Remediation. In K. B. Jensen & R. T. Craig (Eds.), *The international encyclopedia of communication theory and philosophy*. John Wiley & Sons.

Booms, H. (1987). Society and the formation of a documentary heritage: Issues in the appraisal of archival sources. *Archivaria*, *24*(3), 69–107. http://journals.sfu.ca/archivar/index.php/archivaria/article/view/11415/12357

Bourdieu, P. (1977). *Outline of a theory of practice* (Vol. 16). Cambridge University Press.

Bowker, G. C. (2005). *Memory practices in the sciences* (Vol. 205). MIT Press.

boyd, danah, & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679.

Bratton, B. (2016). *The stack*. MIT Press.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101.

Brichford, M. J. (1977). *Archives & manuscripts: Appraisal & accessioning*. Society of American Archivists.

Briet, S. (2006). *What is documentation? : English translation of the classic french*

text. (R. E. Day, L. Martinet, & H. G. B. Anghelescu, Eds.). Scarecrow Press.

Brothman, B. (2001). The past that archives keep: Memory, history, and the preservation of archival records. *Archivaria*, *51*, 48–80.

Brown, R. (1991). Records acquisition strategy and its theoretical foundation: The case for a concept of archival hermeneutics. *Archivaria*, *33*.

Brügger, N. (2016). Digital humanities in the 21st century. *DHQ: Digital Humanities Quarterly*, *10*(2).

Brügger, N., & Schroeder, R. (Eds.). (2017). *The web as history*. UCL Press. http://discovery.ucl.ac.uk/1542998/

Buckland, M. K. (1991). Information as thing. *JASIS*, *42*(5), 351–360. http://skat.ihmc.us/rid=1KR7VC4CQ-SLX5RG-5T39/BUCKLAND(1991)-informationasthing.pdf

Burton, M. (2015). *Blogs as infrastructure for scholarly communication* [PhD thesis]. University of Michigan.

Caswell, M., & Cifor, M. (2016). From human rights to feminist ethics: Radical empathy in the archives. *Archivaria*, *82*, 23–43.

Caswell, M., Punzalan, R., & Sangwand, T.-K. (2017). Critical archival studies: An introduction. *Journal of Critical Library and Information Studies*, *1*(2). https://journals.litwinbooks.com//index.php/jclis/article/view/50

Ceglowski, M. (2011). *Remembrance of links past*. https://blog.pinboard.in/2011/05/remembrance//_of//_links//_past/

Charmaz, K. (2001). Grounded theory. In R. M. Emerson (Ed.), *Contemporary field research: Perspectives and formulatinos* (2nd ed.). Waveland Press.

Christen, K., Merrill, A., & Wynne, M. (2017). A community of relations: Mukurtu

hubs and spokes. *D-Lib Magazine*, *23*(5/6). http://www.dlib.org/dlib/may17/christen/05christen.html

Chun, W. H. K. (2011). *Programmed visions: Software and memory*. MIT Press.

Cifor, M. (2016). Affecting relations: Introducing affect theory to archival discourse. *Archival Science*, *16*(1), 7–31.

Cifor, M., Caswell, M., Migoni, A. A., & Geraci, N. (2018). "What we do crosses over into activism": The politics and practice of community archives. *Acrhives and Public History*, *40*(2), 69–95.

Cifor, M., & Gilliland, A. J. (2016). Affect and the archive, archives and their affects. *Archival Science*, *16*(1), 1–6.

Cobb, J., Pearce-Moses, R., & Surface, T. (2005). ECHO DEPository Project. *Archiving 2005, Final Program and Proceedings*.

Cohn, M. (2016). Convivial decay: Entangled lifetimes in a geriatric infrastructure. *Proceedings of the 19th Acm Conference on Computer-Supported Cooperative Work & Social Computing*, 1509–1521.

Cook, T. (1993). The concept of the archival fonds in the post-custodial era: Theory, problems and solutions. *Archivaria*, *35*, 24–37.

Cook, T. (1994). Electronic records, paper minds: The revolution in information management and archives in the post/custodial and post/modernist era.[Based on a presentation delivered by the author during his november 1993 australian tour.]. *Archives and Manuscripts*, *22*(2), 300–328.

Cook, T. (2004). Macro-appraisal and functional analysis: Documenting governance rather than government 1. *Journal of the Society of Archivists*, *25*(1), 5–18.

Cook, T. (2005). Macroappraisal in theory and practice: Origins, characteristics,

and implementation in canada, 1950–2000. *Archival Science*, *5*(2-4), 101–161.

Cook, T. (2011). We are what we keep; we keep what we are: Archival appraisal past, present and future. *Journal of the Society of Archivists*, *32*(2), 173–189.

Cook, T. (2013). Evidence, memory, identity, and community: Four shifting archival paradigms. *Archival Science*, *13*(2-3), 95–120.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to algorithms* (3rd ed.). MIT Press.

Couture, C. (2005). Archival appraisal: A status report. *Archivaria*, *1*(59), 83–107.

Cox, R. (2000). *Managing records as evidence and information*. Scarecrow Press.

Cox, R. J. (1994). The Documentation Strategy: A different perspective. *Archivaria*, *38*(January), 11–36.

Cox, R. J. (2010). Archivists and collecting. In M. Bates & M. N. Maack (Eds.), *Encyclopedia of library and information sciences* (3rd ed., pp. 208–220). Taylor & Francis.

Cox, R., & Samuels, H. (1988). The archivist's first responsibility: A research agenda to improve the identification and retention of records of enduring value. *The American Archivist*, *51*(1-2), 28–42.

Cubitt, S. (2016). *Finite media: Environmental implications of digital technologies*. Duke University Press.

Deleuze, G., & Guattari, F. (1987). *A thousand plateaus: Capitalism and schizophrenia*. Bloomsbury Publishing.

Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods*. McGraw-Hill.

DHS. (2020). *Test results for disk imaging tool: Roadkil's disk image version 1.6*. Department of Homeland Security.

Diakopoulos, N. (2014). *Algorithmic accountability reporting: On the investigation of black boxes*. Tow Center for Digital Journalism, Columbia University.

Dillon, L., Walker, D., Shapiro, N., Underhill, V., Martenyi, M., Wylie, S., Lave, R., Murphy, M., Brown, P., Data, E., & Initiative, G. (2017). Environmental data justice and the trump administration: Reflections from the environmental data and governance initiative. *Environmental Justice*, *10*(6), 186–192.

DiSalvo, C., Sengers, P., & Brynjarsdóttir, H. (2010). Mapping the landscape of sustainable hci. *Proceedings of the Sigchi Conference on Human Factors in Computing Systems*, 1975–1984.

Dobra, A., & Fienberg, S. E. (2004). How large is the world wide web? In *Web dynamics* (pp. 23–43). Springer.

Dourish, P. (2017). *The stuff of bits: An essay on the materialities of information*. MIT Press.

Dourish, P. (2006). Implications for design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 541–550. http://www.dourish.com/publications/2006/implications-chi2006.pdf

Duncan, S., & Blumenthal, K.-R. (2016). A collaborative model for web archiving ephemeral art resources at new york art resources consortium (nyarc). *Art Libraries Journal*, *41*(2), 116–126.

Duranti, L. (1994). The concept of appraisal and archival theory. *The American Archivist*, 328–344.

Duranti, L. (2010). Concepts and principles for the management of electronic

records, or records management theory is archival diplomatics. *Records Management Journal*, *20*(1), 78–95.

Eastwood, T. (1992). *The archival imagination: Essays in honour of hugh a. Taylor* (B. L. Craig, Ed.; pp. 71–89). Association of Canadian Archivists.

Eastwood, T. (2002). Reflections on the goal of archival appraisal in democratic societies. *Archivaria*, *1*(54). http://journals.sfu.ca/archivar/index.php/archivaria/article/viewFile/12855/14080

Edwards, P., Mayernik, M. S., Batcheller, A., Bowker, G., & Borgman, C. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, 0306312711413314.

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. MIT Press.

Ehn, P. (1988). *Work-oriented design of computer artifacts* [PhD thesis]. Arbetslivscentrum.

Emerson, R. M., Fretz, R. I., & Shaw, L. L. (2011). *Writing ethnographic fieldnotes*. University of Chicago Press.

Ernst, W. (2013). *Digital memory and the archive*. University of Minnesota Press.

Felt, U. (2017). *Handbook of science and technology studies* (U. Felt, R. Fouché, C. A. Miller, & L. Smith-Doerr, Eds.; 4th ed.). MIT Press.

Fenlon, K. S. (2017). *Thematic research collections: Libraries and the evolution of alternative scholarly publishing in the humanities.* [PhD thesis]. University of Illinois Urbana-Champaign.

Fielding, R. (2000). *Architectural styles and the design of network-based software architectures* [PhD thesis, University of California, Irvine]. https://www.ics.uci.edu/\textasciitilde

fielding/pubs/dissertation/top.htm

Fielding, R., Nottingham, M., & Reschke, J. (2014). *Hypertext transfer protocol (http/1.1): Caching* (Nos. 7234). Internet Engineering Task Force. https://tools.ietf.org/html/rfc7234

Fielding, R. T. (1994). Maintaining distributed hypertext infostructures: Welcome to momspider's web. *Computer Networks and ISDN Systems*, *27*(2), 193–204.

Fisher, G. E. (2000). *NSRL data dictionary*. National Institute for Standards; Technology. https://web.archive.org/web/20041015184509/http://www.nsrl.nist.gov/documents/NSRL-DATA-DICTIONARY.pdf

Flinn, A. (2007). Community histories, community archives: Some opportunities and challenges 1. *Journal of the Society of Archivists*, *28*(2), 151–176.

Flinn, A. (2011). Archival activism: Independent and community-led archives, radical public history and the heritage professions. *InterActions: UCLA Journal of Education and Information Studies*, *7*(2).

Flinn, A., Stevens, M., & Shepherd, E. (2009). Whose memories, whose archives? Independent community archives, autonomy and the mainstream. *Archival Science*, *9*(1-2), 71–86.

Foscarini, F. (2017). Archival appraisal in four paradigms. In H. MacNeil & T. Eastwood (Eds.), *Currents of archival thinking* (pp. 107–134). Libraries Unlimited.

Foucault, M. (1982). The subject and power. *Critical Inquiry*, *8*(4), 777–795.

Foucault, M. (1986). Of other spaces. *Diacritics*, *16*(1), 22–27.

Foucault, M. (1991). Governmentality. In *The foucault effect: Studies in governmentality* (pp. 87–104). University of Chicago Press.

Foucault, M. (2012). *Discipline & punish: The birth of the prison*. Vintage.

Foucault, M., Davidson, A. I., & Burchell, G. (2008). *The birth of biopolitics: Lectures at the Collège de France, 1978-1979*. Springer.

Fox-IT. (2017). *Fox-it debunks report on bylock app that landed 75,000 people in jail in turkey*. https://blog.fox-it.com/2017/09/13/fox-it-debunks-report-on-bylock-app-that-landed-75000-people-in-jail-in-turkey/

Galloway, A. R. (2004). *Protocol: How control exists after decentralization*. MIT Press.

Gee, J. (2014). *An introduction to discourse analysis: Theory and method*. Routledge.

Gee, J. P. (2011). *An introduction to critical discourse analysis in education.* (R. Rogers, Ed.; pp. 19–50). Lawrence Erlbaum.

Gee, J. P. (2014). *How to do discourse analysis: A toolkit*. Routledge.

Geertz, C. (1973). *The interpretation of cultures: Selected essays*. Basic books.

Geiger, R. S. (2014). Bots, bespoke, code and the materiality of software platforms. *Information, Communication & Society*, 342–356. http://www.tandfonline.com/doi/full/10.1080/1369118X.2013.873069

Geiger, R. S., & Ribes, D. (2011). Trace ethnography: Following coordination through documentary practices. *44th Hawaii International Conference on System Sciences*, 1–10. http://www.stuartgeiger.com/trace-ethnography-hicss-geiger-ribes.pdf

Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*. University of California Press.

Gillespie, T. (2010). The politics of platforms. *New Media & Society*, *12*(3), 347–364.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gillespie, T., & Seaver, N. (2015). *Critical algorithm studies: A reading list*. https://socialmediacollective.org/reading-lists/critical-algorithm-studies/

Gilliland, A., & McKemmish, S. (2014). The role of participatory archives in furthering human rights, reconciliation and recovery. *Atlanti*, *24*(1), 79–88.

Gitelman, L. (Ed.). (2013). *"Raw data" is an oxymoron.* MIT Press.

Gitelman, L. (2014). *Paper knowledge: Toward a media history of documents*. Duke University Press.

Gokce, Y. (2018). The bylock fallacy: An in-depth analysis of the bylock investigations in turkey. *Digital Investigation*, *26*, 81–91.

Goldman, B. (2018). It's not easy being green(e): Digital preservation in the age of climate change. In *Archival values: Essays in honor of Mark Greene*. Society of American Archivists. https://scholarsphere.psu.edu/concern/generic//_works/bvq27zn11p

Gomes, D., Freitas, S., & Silva, M. J. (2006). Design and selection criteria for a national web archive. In *Research and advanced technology for digital libraries* (pp. 196–207). Springer.

Gossen, G., Demidova, E., & Risse, T. (2015). The iCrawl wizard – supporting interactive focused crawl specification. *Proceedings of the 37th European Conference on Information Retrieval*. http://www.l3s.de/\textasciitilde gossen/publications/gossen_et_al_ecir_2015.pdf

Gracy, K. (2007). Moving image preservation and cultural capital. *Library Trends*, *56*(1), 183–197.

Gracy, K. F. (2004). Documenting communities of practice: Making the case for archival ethnography. *Archival Science*, *4*(3-4), 335–365.

Graham, M. (n.d.). *Cloudflare and the wayback machine, joining forces for a more reliable web*. Internet Archive Blogs. http://blog.archive.org/2020/09/17/internet-archive-partners-with-cloudflare-to-help-make-the-web-more-useful-and-reliable/

Graham, S., & Thrift, N. (2007). Out of order understanding repair and maintenance. *Theory, Culture & Society*, *24*(3), 1–25.

Greene, M. (1998). "The surest proof": A utilitarian approach to appraisal. *Archivaria*, 127–169.

Gusterson, H. (1997). Studying up revisited. *PoLAR: Political and Legal Anthropology Review*, *20*(1), 114–119.

Ham, F. (1984). Archival choices: Managing the historical record in an age of abundance. *The American Archivist*, *47*(1), 11–22.

Ham, F. G. (1975). The archival edge. *The American Archivist*, *38*(1), 5–13. http://americanarchivist.org/doi/pdf/10.17723/aarc.38.1.7400r86481128424

Ham, F. G. (1981). Archival strategies for the post-custodial era. *The American Archivist*, *44*(3), 207–216.

Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, *14*(3), 575–599.

Harris, V. (1998). Postmodernism and archival appraisal: Seven theses. *South African Archives Journal*, *40*, 48–50.

Harris, V. (2002). The archival sliver: Power, memory, and archives in South Africa. *Archival Science*, *2*(1-2), 63–86.

Harvey, R., & Thompson, D. (2010). Automating the appraisal of digital materials. *Library Hi Tech*, *28*(2), 313–322.

Hedstrom, M. (1991). Understanding electronic incunabula: A framework for research on electronic records. *The American Archivist*, *54*(3), 334–354.

Hedstrom, M. (2002). Archives, memory, and interfaces with the past. *Archival Science*, *2*(1-2), 21–43.

Hennessey, J., & Ge, S. X. (2013). A cross disciplinary study of link decay and the effectiveness of mitigation techniques. *BMC Bioinformatics*, *14*(14).

Hitchcock, S., Brody, T., & Hey, J. M. N. (2007). Digital preservation service provider models for institutional repositories. *D-Lib Magazine*, *13*(5/6). http://www.dlib.org/dlib/may07/hitchcock/05hitchcock.html

Holstein, J. A., & Gubrium, J. F. (2011). Animating interview narratives. *Qualitative Research*, *3*, 149–167.

Hoskins, A. (2018). Memory of the multitude: The end of collective memory. In A. Hoskins (Ed.), *Digital memory studies: Media pasts in transition* (pp. 85–109). Routledge.

Hu, T.-H. (2015). *A prehistory of the cloud*. MIT Press.

Huvila, I. (2008). Participatory archive: Towards decentralised curation, radical user orientation, and broader contextualisation of records management. *Archival Science*, *8*(1), 15–36.

Huvila, I. (2015). The unbearable lightness of participating? Revisiting the discourses of 'participation' in archival literature. *Journal of Documentation*, *43*, 29–41.

Ivanov, A. O. (2017). Practice theory: A new approach for archival and recordkeep-

ing research. *Records Management Journal*, *27*(2), 104–124.

Jackson, S. J. (2014). *Media technologies: Essays on communication, materiality and society* (P. Boczkowski & K. Foot, Eds.; pp. 221–239). MIT Press. http://sjackson.infosci.cornell.edu/RethinkingRepairPROOFS(reduced)Aug2013.pdf

Jackson, S. J., & Buyuktur, A. (2014). Who killed WATERS? Mess, method, and forensic explanation in the making and unmaking of large-scale science networks. *Science, Technology, & Human Values*, *39*(2), 285–308.

Jackson, S. J., Gillespie, T., & Payette, S. (2014). The policy knot: Re-integrating policy, practice and design in CSCW studies of social computing. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 588–602.

Jackson, S. J., & Kang, L. (2014). *Breakdown, obsolescence and reuse: HCI and the art of repair*. http://sjackson.infosci.cornell.edu/Jackson//&Kang//_BreakdownObsolescenceReuse(CHI2014).pdf

Jacobs, I., & Walsh, N. (2004). *Architecture of the world wide web, volume one*. World Wide Web Consortium. http://www.w3.org/TR/webarch/

Jacobsen, T., Punzalan, R. L., & Hedstrom, M. L. (2013). Invoking collective memory: Mapping the emergence of a concept in archival science. *Archival Science*, *13*(2-3), 217–251.

Janée, G., Frew, J., & Moore, T. (2009). Relay-supporting archives: Requirements and progress. *International Journal of Digital Curation*, *4*(1), 57–70. http://www.ijdc.net/index.php/ijdc/article/view/102

Jasanoff, S. (2006). *States of knowledge: The co-production of science and the social order*. Routledge.

Jenkinson, H. (1922). *A manual of archive administration including the problems of war archives and archive making*. Clarendon Press.

Jimerson, R. (2013). Archivists and social responsibility: A response to mark greene. *The American Archivist*, *76*(2), 335–345.

Jimerson, R. C. (2009). *Archives power: Memory, accountability, and social justice*. Society of American Archivists.

Kahle, B. (2007). Universal access to all knowledge. *The American Archivist*, *70*(1), 23–31.

Karasti, H., & Blomberg, J. (2018). Studying infrastructuring ethnographically. *Computer Supported Cooperative Work (CSCW)*, 1–33.

Kelty, C. M. (2008). *Two bits: The cultural significance of free software*. Duke University Press. http://twobits.net/

Kensing, F., & Blomberg, J. (1998). Participatory design: Issues and concerns. *Computer Supported Cooperative Work (CSCW)*, *7*(3-4), 167–185.

Ketelaar, E. (2001). Tacit narratives: The meanings of archives. *Archival Science*, *1*(2), 131–141.

Ketelaar, E. (2005). Recordkeeping and societal power. In *Archives: Recordkeeping in society.* Charles Stuart University.

Kirschenbaum, M. G. (2008). *Mechanisms: New media and the forensic imagination*. MIT Press.

Kitchin, R. (2016). Thinking critically about and researching algorithms. *Information, Communication & Society*, *20*(1), 1–16.

Kitchin, R., & Lauriault, T. P. (2014). *Towards critical data studies: Charting and unpacking data assemblages and their work* (The Programmable City Work-

ing Paper 2). The Programmable City; The Programmable City Working Paper 2. https://papers.ssrn.com/sol3/papers.cfm?abstract//_id=2474112

Kittler, F. A. (1999). *Gramophone, film, typewriter*. Stanford University Press.

Klein, M., Balakireva, L., & Van de Sompel, H. (2018). Focused crawl of web archives to build event collections. *arXiv Preprint arXiv:1804.01603*.

Knuth, D. (1998). *The art of computer programming: Volume 3, sorting and searching* (2nd, Ed.; Vol. 3). Addison-Wesley.

Kosnik, A. D. (2016). *Rogue archives: Digital cultural memory and media fandom*. MIT Press.

Kranzberg, M. (1986). Technology and history:" Kranzberg's laws". *Technology and Culture*, *27*(3), 544–560.

Kuny, T. (1997). A digital dark ages? Challenges in the preservation of electronic information. *Proceedings of the 63rd International Federation of Library Associations*. https://archive.ifla.org/IV/ifla63/63kuny1.pdf

Kunze, J., LIttman, J., Madden, E., Scancella, J., & Adams, C. (2018). *The bagit file packaging format (v1.0)* (RFC 8493). Internet Engineering Task Force.

Kuutti, K., & Bannon, L. J. (2014). The turn to practice in HCI: Towards a research agenda. *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, 3543–3552. http://dl.acm.org/citation.cfm?id=2557111

Lampland, M., & Star, S. L. (2009). *Standards and their stories: How quantifying, classifying, and formalizing practices shape everyday life*. Cornell University Press.

Lasfrgues, F., Oury, C., & Wendland, B. (2008). Legal deposit of the french web: Harvesting strategies for a national domain. *Proceedings from IWAW '08: 8th International Workshop for Web Archiving*.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Harvard University Press.

Latour, B. (1990). Technology is society made durable. *The Sociological Review*, *38*, 103–131.

Latour, B. (1999). *Pandora's hope: Essays on the reality of science studies*. Harvard University Press.

Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford University Press.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.

Law, J. (2009). *The new Blackwell companion to social theory* (B. S. Turner, Ed.; pp. 141–158). Wiley-Blackwell.

Law, J., & Lien, M. E. (2012). Slippery: Field notes in empirical ontology. *Social Studies of Science*, *43*(3), 363–378.

Lee, C. (2018). Computer-assisted appraisal and selection of archival materials. *2018 IEEE International Conference on Big Data*.

Lee, C. A. (Ed.). (2011). I, digital: Personal collections in the digital era. In *I, Digital: Personal Collections in the Digital Era. Chicago, IL: Society of American Archivists*. Society of American Archivists.

Lee, C. A., & Tibbo, H. (2011). Where's the archivist in digital curation? Exploring the possibilities through a matrix of knowledge and skills. *Archivaria*, *72*(Fall).

Lemke, T. (2019). *Foucault's analysis of modern governmentality: A critique of political reason* (E. Butler, Trans.). Verso.

Lepore, J. (2015). The cobweb: Can the internet be archived? *The New Yorker*.

http://www.newyorker.com/magazine/2015/01/26/cobweb

Lessig, L. (2006). *Code: Version 2.0*. Basic Books.

Levy, D. (2001). *Scrolling forward*. Arcade.

Linde, C. (2009). *Working the past: Narrative and institutional memory*. Oxford University Press.

Lyle, D. P. (2017). *Vintage computer games, missing aircraft, and an amazing forensic resource: An interview with douglas white of the national institute of science and technology's national software reference library (nsrl)*. https://www.blogtalkradio.com/suspensemagazine/2017/02/04/crime-and-science-radio-with-special-guest-douglas-white

Lyle, J. (2002). NSRL project. *Regional Computer Forensices Group GMU 2002*. https://www.nist.gov/system/files/nsrl//_gmu2002.pdf

Mackenzie, A. (2017). *Machine learners: Archaeology of a data practice*. MIT Press.

Madhavan, J., Ko, D., Kot, Ł., Ganapathy, V., Rasmussen, A., & Halevy, A. (2008). Google's deep web crawl. *Proceedings of the VLDB Endowment*, *1*(2), 1241–1252.

Maemura, E., Worby, N., Milligan, I., & Becker, C. (2018). If these crawls could talk: Studying and documenting web archives provenance. *Journal of the Association for Information Science and Technology*, *69*(10), 1223–1233. https://tspace.library.utoronto.ca/handle/1807/82840

Maestri, L., & Wakkary, R. (2011). Understanding repair as a creative process of everyday design. *Proceedings of the 8th Acm Conference on Creativity and Cognition*, 81–90.

Manovich, L. (1999). Manovich: Database as a symbolic form. *Convergence: The*

International Journal of Research into New Media Technologies.

Masanès, J. (2006a). Selection for web archives. In J. Masanès (Ed.), *Web archiving issues and methods*. Springer.

Masanès, J. (Ed.). (2006b). *Web archiving: Issues and methods*. Springer.

Masanès, J. (2006c). Web archiving methods and approaches: A comparative study. *Library Trends*, *54*(1), 72–90.

Mayer-Schönberger, V. (2011). *Delete: The virtue of forgetting in the digital age*. Princeton University Press.

McDonough, J., Olendorf, R., Kirschenbaum, M., Kraus, K., Reside, D., Donahue, R., Phelps, A., Egert, C., Lowood, H., & Rojo, S. (2010). *Preserving virtual worlds final report*. University of Illinois at Urbana-Champaign; Citeseer.

McKemmish, S., & others. (1996). Evidence of me. *Archives and Manuscripts*, *24*(1), 28.

McKemmish, S., Upward, F., & Reed, B. (2010). Records continuum model. In M. Bates & M. N. Maack (Eds.), *Encyclopedia of library and information sciences*. Taylor
& Francis.

Meng, A., & DiSalvo, C. (2018). Grassroots resource mobilization through counter-data action. *Big Data & Society*, *5*(2), 2053951718796862.

Milligan, I., Ruest, N., & Lin, J. (2016). Content selection and curation for web archiving: The gatekeepers vs. The masses. *Proceedings of the Joint Conference on Digital Libraries*. https://cs.uwaterloo.ca/\textasciitilde jimmylin/publications/Milligan_etal_JCDL2016.pdf

Mohr, G., Stack, M., Rnitovic, I., Avery, D., & Kimpton, M. (2004). Introduction

to heritrix. *4th International Web Archiving Workshop*. https://webarchive.jira.com/wiki/download/attachments/5441/Mohr-et-al-2004.pdf

Mol, A. (2002). *The body multiple: Ontology in medical practice*. Duke University Press.

Muller, S., Feith, J. A., & Fruin, R. (1940). *Manual for the arrangement and description of archives*. The H. W. Wilson Company.

Nader, L. (1972). Up the anthropologist: Perspectives gained from studying up. In *Reinventing anthropology*. Pantheon.

Neal, K. M. (2002). Cultivating diversity: The donor connection. *Collection Management*, *27*(2), 33–42.

Nicolini, D. (2009). Zooming in and out: Studying practices by switching theoretical lenses and trailing connections. *Organization Studies*, *30*(12), 1391–1418.

NIST. (2001). *Information technology laboratory technical accomplishments: Enabling a better future* (NISTIR 6815). National Institute of Standards; Technology. https://www.govinfo.gov/content/pkg/GOVPUB-C13-5f3d65558bdc85b1a14d8a0d99afb51f/pdf/GOVPUB-C13-5f3d65558bdc85b1a14d8a0d99afb51f.pdf

NIST. (2002a). *Data formats of the nsrl reference data set (rds) distribution*. https://web.archive.org/web/20040808023938/http://www.nsrl.nist.gov/documents/Data-Formats-of-the-NSRL-Reference-Data-Set-12.pdf

NIST. (2002b). *Information technology laboratory technical accomplishments: Enabling a better future* (NISTIR 6909). National Institute of Standards; Technology. https://www.govinfo.gov/content/pkg/GOVPUB-C13-a36bbdfba9bacc82fb508fa20e4bfb39/pdf/GOVPUB-C13-a36bbdfba9bacc82fb508fa20e4bfb39.pdf

NIST. (2003). *National software reference library (nsrl) project web site*. https:

//web.archive.org/web/20030320060504/http://www.nsrl.nist.gov:80/

NIST. (2004). *NSRL and recent cryptographic news*. https://web.archive.org/web/20161107085243/http://www.nsrl.nist.gov/collision.html

NIST. (2017). *NSRL and recent cryptographic news*. https://web.archive.org/web/20170625051653/https://www.nsrl.nist.gov/collisionSHA1.html

NIST. (2018). *Database of software "fingerprints" expands to include computer games*. https://www.nist.gov/news-events/news/2018/09/database-software-fingerprints-expands-include-computer-games

NIST. (2020a). *National software reference library (nsrl)*. https://www.nist.gov/itl/ssd/software-quality-group/national-software-reference-library-nsrl

NIST. (2020b). *NIST special database 28*. National Institute of Standards; Technology. https://www.nist.gov/srd/nist-special-database-28

Nwala, A. C., Weigle, M. C., & Nelson, M. L. (2018). Bootstrapping web archive collections from social media. *Proceedings of the 29th Conference on Hypertext and Social Media*.

Ogden, J. (2019). *Saving the web: Facets of web archiving in everyday practice* [PhD thesis]. University of Southampton.

Ogden, J., Halford, S., & Carr, L. (2017). Observing web archives. *Proceedings of WebSci'17*. https://eprints.soton.ac.uk/410123/

O'Toole, J. M. (2004). Back to the future: Ernst posner's archives in the ancient world. *American Archivist*, *67*(Fall/Winter), 161–175.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web.* Stanford Infolab.

Passi, S., & Jackson, S. (2017). Data vision: Learning to see through algorithmic

abstraction. *Proceedings of the 2017 ACM Conference on Computer Supported Co-operative Work and Social Computing*, 2436–2447.

Pearce-Moses, R., & Baty, L. A. (2005). *A glossary of archival and records terminology*. Society of American Archivists Chicago, IL. https://www2.archivists.org/glossary

Pearce-Moses, R., & Kaczmarek, J. (2005). An arizona model for preservation and access of web documents. *DTTP: Documents to the People*, *33*(1), 17–24.

Pinch, T. J. (1987). *Social construction of technological systems* (W. E. Bijke & T. P. Hughes, Eds.). MIT Press.

Posner, E. (1972). *Archives in the ancient world*. Harvard University Press.

Postill, J. (2010). *Theorising media and practice* (B. Bräuchler & J. Postill, Eds.). Berghahn Books.

Protocol Labs. (2020). *IPFS powers the distributed web*. https://ipfs.io

Punzalan, R. (2014). Understanding virtual reunification. *The Library Quarterly: Information, Community, Policy*, *84*(3), 294–323.

Punzalan, R. L. (2009). 'All the things we cannot articulate': Colonial leprosy archives and community commemoration. In J. A. Bastian & B. Alexander (Eds.), *Community archives: The shaping of memory*. Facet Publishing.

Punzalan, R. L., & Caswell, M. (2016). Critical directions for archival approaches to social justice. *Library Quarterly*, *86*(1), 25–42.

Rapport, L. (1981). No grandfather clause: Reappraising accessioned records. *The American Archivist*, *44*(2), 143–150.

Ribes, D., & Finholt, T. (2009). The long now of technology infrastructure: Articulating tensions in development. *Journal of the Association for Information Systems*,

10(5).

Risse, T., Dietze, S., Peters, W., Doka, K., Stavrakas, Y., & Senellart, P. (2012). Exploiting the social and semantic web for guided web archiving. In *Theory and practice of digital libraries* (pp. 426–432). Springer.

Rollason-Cass, S., & Reed, S. (2015). Living movements, living archives: Selecting and archiving web content during times of social unrest. *New Review of Information Networking*, *20*(1-2), 241–247.

Rose, N. (1999). *Governing the soul: The shaping of the private self*. Free Association Books.

Rosenthal, D. (2012). *Let's just keep everything forever in the cloud.* http://blog.dshr.org/2012/05/lets-just-keep-everything-forever-in.html

Rosner, D. K., & Ames, M. (2014). Designing for repair?: Infrastructures and materialities of breakdown. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 319–331. http://people.ischool.berkeley.edu/\textasciitilde daniela/files/cscw14-rosner-repair.pdf

Rothenberg, J. (1999). *Avoiding technological quicksand: Finding a viable technical foundation fo rDigital preservation* (Nos. ED426715). Council on Library; Information Resources.

Russell, A. L. (2014). *Open standards and the digital age*. Cambridge University Press.

Russell, A. L., & Vinsel, L. (2018). After innovation, turn to maintenance. *Technology and Culture*, *59*(1).

Ryle, G. (1968). The thinking of thoughts. *Studies in Anthropology*, *11*(11).

Salaheldeen, H. M., & Nelson, M. L. (2013). Resurrecting my revolution. In *Research and advanced technology for digital libraries* (pp. 333–345). Springer.

SalahEldeen, H. M., & Nelson, M. L. (2013). Carbon dating the web: Estimating the age of web resources. *Proceedings of the 22nd International Conference on World Wide Web Companion*, 1075–1082. http://arxiv.org/abs/1304.5213

Samar, T., Huurdeman, H. C., Ben-David, A., Kamps, J., & Vries, A. de. (2014). Uncovering the unarchived web. *Proceedings of the 37th International Acm Sigir Conference on Research & Development in Information Retrieval*, 1199–1202. http://humanities.uva.nl/\textasciitilde kamps/publications/2014/sama:unco14.pdf

Samuels, H. W. (1986). Who controls the past. *The American Archivist*, 109–124. http://americanarchivist.org/doi/abs/10.17723/aarc.49.2.t76m2130txw40746

Samuels, H. W. (1991). Improving our disposition: Documentation strategy. *Archivaria*, *33*(1), 125–140.

Sanderson, R., Phillips, M., & Sompel, H. V. de. (2011). *Analyzing the persistence of referenced web resources with Memento*. Open Repositories 2011 Conference. http://arxiv.org/abs/1105.3459

Schafer, V., Musiani, F., & Borelli, M. (2016). Negotiating the web of the past: Web archiving, governance and sts. *French Journal for Media Research*, *6*.

Schellenberg, T. R. (1956). *Modern archives: Principles and techniques*. University of Chicago Press. http://catalog.hathitrust.org/Record/003147122

Schlanger, Z. (2017). *Rogue scientists race to save climate data from trump*. https://www.wired.com/2017/01/rogue-scientists-race-save-climate-data-trump/

Schneider, S. M., Foot, K., Kimpton, M., & Jones, G. (2003). Building thematic web collections: Challenges and experiences from the september 11 web archive

and the election 2002 web archive. *3rd Ecdl Workshop on Web Archives*.

Schwartz, J. M. (1995). ”We make our tools and our tools make us”: Lessons from photographs for the practice, politics, and poetics of diplomatics. *Archivaria*, *40*(Fall), 40–74.

Schwartz, J. M., & Cook, T. (2002). Archives, records and power: The making of modern memory. *Archival Science*, *2*, 1–19.

Scott, J. (2019). *The hitchhiker's guide to the galaxy (solid-gold edition) source code collection*. https://github.com/historicalsource/hitchhikersguide-gold

Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. Yale University Press.

Scott, J. C. (2017). *Against the grain: A deep history of the earliest states.* Yale University Press.

Seaver, N. (2013). *Knowing algorithms*. http://nickseaver.net/s/seaverMiT8.pdf

Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, *4*(2).

Sheffield, R. T. (2018). Facebook live as record making technology. *Archivaria*, *85*.

Star, S. L. (1990). Power, technology and the phenomenology of conventions: On being allergic to onions. *The Sociological Review*, *38*(S1), 26–56.

Star, S. L. (1999). The ethnography of infrastructure. *American Behavioral Scientist*, *43*(3), 377–391.

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, *7*(1), 111–134.

Starosielski, N. (2015). *The undersea network*. Duke University Press.

Stevens, H. (2018). Hans peter luhn and the birth of the hashing algorithm. In *IEEE Spectrum*. https://spectrum.ieee.org/tech-history/silicon-revolution/hans-peter-luhn-and-the-birth-of-the-hashing-algorithm

Suchman, L. (1985). *Plans and situated actions: The problem of human-machine communication.* Xerox Corporation.

Taylor, H. (1988). My very act and deed: Some reflections on the role of textual records in the conduct of affairs. *American Archivist*, *51*(Fall), 456–469.

Taylor, H. A. (1984). Information ecology and the archives of the 1980s. *Archivaria*, *18*, 25–37.

Taylor, H. A. (1992). *The archival imagination: Essays in honour of Hugh A. Taylor* (B. L. Craig, Ed.). Association of Canadian Archivists.

Taylor, L. (2017). What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society*, *4*(2), 2053951717736335.

The WARC format 1.1 (ISO 28500:2017). (2017). International Organization for Standardization. https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/

Trace, C. B. (2010). On or off the record? Notions of value in the archive. *Currents of Archival Thinking*, 47–68.

Tschan, R. (2002). A comparison of Jenkinson and Schellenberg on appraisal. *The American Archivist*, *65*(2), 176–195.

Tufekci, Z. (2017). *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.

University, S. (2000). *Guide to the stephen m. Cabrinety collection in the history*

of microcomputing, ca. 1975-1995. Online Archive of California. https://oac.cdlib.org/findaid/ark:/13030/kt529018f2/

Upward, F. (1996). Structuring the records continuum (series of two parts) part 1: Post custodial principles and properties. *Archives and Manuscripts*, *24*(2), 268.

Upward, F. (1997). Structuring the records continuum (series of two parts) part 2: Structuration theory and recordkeeping. *Archives and Manuscripts*, *25*(1), 10.

Vertesi, J. (2015). *Seeing like a rover*. University of Chicago Press.

Walford, A. (2017). Raw data: Making relations matter. *Social Analysis*, *61*(2), 65–80.

Weinberg, J., & Zimmermann, A. (n.d.). *Philosophers on gpt-3*. Daily Nous. https://dailynous.com/2020/07/30/philosophers-gpt-3/#chalmers

Welch, I., Rehfeld, N., Cochrane, E., & Suchodoletz, D. von. (2012). A practical approach to system preservation workflows. *PIK - Praxis Der Informationsverarbeitung Und Kommunikation*, *35*(4), 269–280.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge University Press.

Winner, L. (1980). Do artifacts have politics? *Daedalus*, 121–136.

Wittgenstein, L. (1953). *Philosophical investigations*. Prentice Hall.

Wodak, R., & Meyer, M. (Eds.). (2001). *Methods of critical discourse analysis*. Sage.

Yakel, E. (2003). Archival representation. *Archival Science*, *3*, 1–25.

Yakel, E. (2007). Digital curation. *OCLC Systems & Services: International Digital Library Perspectives*, *23*(4), 335–340.

Yang, S., Chitturi, K., Wilson, G., Magdy, M., & Fox, E. A. (2012). A study of automation from seed URL generation to focused web archive development: The CTRnet context. *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, 341–342.

Yeo, G. (2007). Concepts of record (1): Evidence, information, and persistent representations. *The American Archivist*, *70*(2), 315–343.

Yeo, G. (2008). Concepts of record (2): Prototypes and boundary objects. *The American Archivist*, *71*(1), 118–143.

Young, J. M. (1985). Annotated bibliography on appraisal. *The American Archivist*.

Zinn, H. (1977). Secrecy, archives, and the public interest. *The Midwestern Archivist*, *2*(2), 14–26.

Zittrain, J., Albert, K., & Lessig, L. (2014). Perma: Scoping and addressing the problem of link and reference rot in legal citations. *Legal Information Management*, *14*(02), 88–99.

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, *30*(1), 75–89.